



ESCOLA NAVAL

talant de bi-faire



Maria Inês Neves de Sousa

*Data mining for anomaly detection in maritime
traffic data*

**Dissertação para obtenção do grau de Mestre em Ciências
Militares Navais, na especialidade de Marinha**



**Alfeite
2018**



ESCOLA NAVAL

talant de bi-faire



Maria Inês Neves de Sousa

Data mining for anomaly detection in maritime traffic data

**Dissertação para obtenção do grau de Mestre em Ciências Militares
Navais, na especialidade de Marinha**

Orientação de: CFR RES Aldino Manuel dos Santos de Campos

Coorientação de: Victor José de Almeida e Sousa Lobo e de
Rui Filipe Pedroso Maia

O aluno mestrando,

O orientador,

ASPOF Neves de Sousa

CFR RES Aldino Campos

**Alfeite
2018**



"We are drowning in information, but starving for knowledge."

John Naisbitt



Data mining for anomaly detection in maritime traffic data



To my family...

Thank you.



Data mining for anomaly detection in maritime traffic data

Acknowledgements

There are many people I would like to thank.

To my Thesis Supervisor, CFR RES Aldino Campos, for his guidance and prompt response. His patience and composure were very important, helping me to stay focused and motivated during the course of this dissertation.

To Rui Maia, for his enthusiasm and all the help given throughout this year, providing me with the tools that I needed to choose the right direction and complete my dissertation with success.

To Professor Victor Lobo, for his insightful comments and help given, his advices were of extreme importance.

To CFR Fidalgo Neves, for all the interest demonstrated and willingness to help whenever he could, especially in the early stages of this dissertation.

I would also like to thank all the Organisations and participants that helped me throughout this project, particularly *Direção de Tecnologias de Informação e Comunicações*, *INOV-INESC*, *Centro de Operações Marítimas*, *Direção de Análise e Gestão da Informação* and *Instituto Hidrográfico*.

To my family, for all the love, support and encouragement given, it was essential.

To Alexandre, for his patience and for making this journey easier. It would not be the same without you.



Data mining for anomaly detection in maritime traffic data



Abstract

For the past few years, oceans have become once again, an important means of communication and transport. In fact, traffic density throughout the globe has suffered a substantial growth, which has risen some concerns. With this expansion, the need to achieve a high Maritime Situational Awareness (MSA) is imperative. At the present time, this need may be more easily fulfilled thanks to the vast amount of data available regarding maritime traffic. However, this brings in another issue: data overload. Currently, there are so many data sources, so many data to obtain information from, that the operators cannot handle it. There is a pressing need for systems that help to sift through all the data, analysing and correlating, helping in this way the decision making process.

In this dissertation, the main goal is to use different sources of data in order to detect anomalies and contribute to a clear Recognised Maritime Picture (RMP). In order to do so, it is necessary to know what types of data exist and which ones are available for further analysis. The data chosen for this dissertation was Automatic Identification System (AIS) and *Monitorização Contínua das Atividades da Pesca* (MONICAP) data, also known as Vessel Monitoring System (VMS) data. In order to store 1 year worth of AIS and MONICAP data, a PostgreSQL database was created. To analyse and draw conclusions from the data, a data mining tool was used, namely, Orange. Tests were conducted in order to assess the correlation between data sources and find anomalies.

The importance of data correlation has never been so important and with this dissertation the aim is to show that there is a simple and effective way to get answers from great amounts of data.

Keywords: maritime situational awareness, data, maritime traffic, AIS, MONICAP.



Data mining for anomaly detection in maritime traffic data

Resumo

Nos últimos anos, os oceanos tornaram-se, mais uma vez, um importante meio de comunicação e transporte. De facto, a densidade de tráfego global sofreu um crescimento substancial, o que levantou algumas preocupações. Com esta expansão, a necessidade de atingir um elevado Conhecimento Situacional Marítimo (CSM) é imperativa. Hoje em dia, esta necessidade pode ser satisfeita mais facilmente graças à vasta quantidade de dados disponíveis de tráfego marítimo. No entanto, isso leva a outra questão: sobrecarga de dados. Atualmente existem tantas fontes de dados, tantos dados dos quais extrair informação, que os operadores não conseguem acompanhar. Existe uma necessidade premente para sistemas que ajudem a escrutinar todos os dados, analisando e correlacionando, contribuindo desta maneira ao processo de tomada de decisão.

Nesta dissertação, o principal objetivo é usar diferentes fontes de dados para detetar anomalias e contribuir para uma clara *Recognised Maritime Picture* (RMP). Para tal, é necessário saber que tipos de dados existem e quais é que se encontram disponíveis para análise posterior. Os dados escolhidos para esta dissertação foram dados *Automatic Identification System* (AIS) e dados de Monitorização Contínua das Atividades da Pesca (MONICAP), também conhecidos como dados de *Vessel Monitoring System* (VMS). De forma a armazenar dados correspondentes a um ano de AIS e MONICAP, foi criada uma base de dados em PostgreSQL. Para analisar e retirar conclusões, foi utilizada uma ferramenta de data mining, nomeadamente, o Orange. De modo a que pudesse ser avaliada a correlação entre fontes de dados e serem detetadas anomalias foram realizados vários testes.

A correlação de dados nunca foi tão importante e pretende-se com esta dissertação mostrar que existe uma forma simples e eficaz de obter respostas de grandes quantidades de dados.

Palavras-chave: conhecimento situacional marítimo, dados, tráfego marítimo, AIS, MONICAP.



Data mining for anomaly detection in maritime traffic data



Contents

ACKNOWLEDGEMENTS	IX
ABSTRACT	XI
RESUMO	XIII
CONTENTS	XV
LIST OF FIGURES	XVII
LIST OF TABLES	XXI
LIST OF ABBREVIATIONS AND ACRONYMS	XXIII
1. INTRODUCTION	2
1.1. Motivation	7
1.2. Topic's relevance	7
1.3. Goals	8
1.4. Structure	8
2. LITERATURE REVIEW	10
2.1. Maritime Situational Awareness	10
2.2. Maritime Domain Data	16
2.3. Concept of Anomaly	28
2.4. Anomaly Detection	29
2.5. Data mining tools	30
3. WORKFLOW AND DATA PROCESSING	38
3.1. Workflow	38
3.2. Data collection	39
3.3. Data Processing: AIS and MONICAP	39
3.4. PostgreSQL Database	44
4. DATA ANALYSIS TOOLS	50
4.1. Selected area	50
4.2. Data Mining for anomaly detection using Orange	51
5. TESTS AND RESULTS	58
5.1. Data correlation tests and results	58
5.2. Anomaly Detection tests and results	72
6. CONCLUSION	86
6.1. Summary	86
6.2. Constraints	87
6.3. Future work	87
REFERENCES	89
APPENDIX A	95
ANNEX A – INTERNATIONAL REQUIREMENTS FOR AIS CARRIAGE	97
ANNEX B - RULE 10: TRAFFIC SEPARATION SCHEMES	99



Data mining for anomaly detection in maritime traffic data

List of figures

Figure 1 - Main international maritime routes (Jornal da Economia do Mar, n.d.).....	3
Figure 2 - Real time maritime traffic (Marine Traffic, n.d.).	3
Figure 3 - Data, information, knowledge and wisdom (context vs understanding) (Marinha, 2018).	7
Figure 4 - Challenges of Maritime Situational Awareness	14
Figure 5 –AIS Functioning – Slots (Navigation Center, n.d.).	19
Figure 6 – Example of information given by AIS.....	21
Figure 7 - AIS data, reports inside the Maritime Operational Area.....	24
Figure 8 - MONICAP system's functioning (Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos, n.d.).	25
Figure 9 - Continuous monitoring equipment (XSEALENCE, n.d.).	26
Figure 10 - Example of MONICAP messages.	26
Figure 11 - Collective Anomaly example (Chandola, 2009).....	29
Figure 12 - Knowledge discovery process.....	31
Figure 13 - Rapid Miner environment.	32
Figure 14 - WEKA environment.	33
Figure 15 - Weka GUI Chooser.....	33
Figure 16 - Orange environment.	34
Figure 17 - Orange main widget groups.....	35
Figure 18 - Steps of the research.....	38
Figure 19 - Excel daily file.	40
Figure 20 - Conversion from .xlsx files to .csv.	41
Figure 21 - Aggregation of the daily .csv files into monthly .csv files.	41
Figure 22 - Shapefile created through ArcGIS.....	42
Figure 23 - QGIS vector layer added.	43
Figure 24 – QGIS to PostgreSQL connection.	43
Figure 25 – Database in the pgAdmin interface.	44
Figure 26 - pgAdmin tables of AIS_MONICAP_2017 database.....	46
Figure 27 - Example of one of the AIS tables created.	46
Figure 28 - Selected area (Daftlogic, n.d.).....	50
Figure 29 - Traffic density in the chosen area (Marine Traffic, n.d.).....	51
Figure 30 - Orange connection to PostgreSQL database.....	52

Figure 31 - Distance calculation methods available in Orange.	53
Figure 32 - Orange's impute widget.	53
Figure 33 - Different classification based on k value.	54
Figure 34 - Example of Gaussian distribution (Boost, n.d.).....	55
Figure 35 - Orange's outlier widget.	56
Figure 36 - Data file insertion.	58
Figure 37 - Selection of features.	59
Figure 38 - Distance selected, Euclidean.	59
Figure 39 - Distance matrix application.	60
Figure 40 – Example of color gradient based on distances.	61
Figure 41 - Distance matrix results between AIS data.....	62
Figure 42 - Distance matrix results.....	62
Figure 43 - Values used to analyse results of distance matrix.....	63
Figure 44 - Distance obtained from correlating AIS and MONICAP data.....	64
Figure 45 – Values used to analyse distance matrix with 8 minutes difference.	65
Figure 46 - Results obtained from correlating AIS and MONICAP data.....	66
Figure 47 - Orange scheme without time data.....	66
Figure 48 - Distance matrix results without time data.....	67
Figure 49 - Distance matrix results for test 2.....	68
Figure 50 - Distance matrix results in test 3	69
Figure 51 - AIS and MONICAP data used on test 3.	70
Figure 52 - Results after Impute widget.....	71
Figure 53 - Distance matrix results in test 3 b).	72
Figure 54 - Anomaly detection test Orange Workflow	73
Figure 55 - Impute widget erroneous results	74
Figure 56 - Branch of the workflow - Analysing records SOG = 0.....	74
Figure 57 - Records with SOG = 0 on 1 st of March, corresponding to a total of 3024 records.....	75
Figure 58 - Outlier widget configuration	75
Figure 59 - a) Outlier Records; b) Inlier Records.....	76
Figure 60 - Vessel with MMSI 255125111 route on 1 st of March.	77
Figure 61 - SOG Outlier detection.....	77
Figure 62 - Results of SOG outlier detection.....	78
Figure 63 - SOG outlier detection after filtering.	78
Figure 64 - Traffic separation scheme branch.	79



Figure 65 - Features selection for traffic separation scheme.	80
Figure 66 - Outlier widget for traffic separation scheme.	81
Figure 67 - TSS of Cape Roca.....	81
Figure 68 - Crossing a traffic lanes according to International Regulations for Preventing Collisions at Sea.	82
Figure 69 - Traffic separation scheme, a) Every record; b) Outlier records.	83
Figure 70 - TSS anomalous records.	83



Data mining for anomaly detection in maritime traffic data



List of tables

Table 1 – Example of MONICAP data format.....	26
Table 2 – AIS and VMS comparison (Navigation Center, n.d.).....	28
Table 3 - Summary table of the 3 data mining tools explored (Predictive Analytics Today, n.d.).	36
Table 4 - PostgreSQL processing capability (Database guide, n.d.).....	44
Table 5 - Summary of the conducted tests.....	84



Data mining for anomaly detection in maritime traffic data

List of Abbreviations and Acronyms

ACID	Atomic, Consistent, Isolation and Durability
AIS	Automatic Identification System
AMT	<i>Autoridade da Mobilidade e dos Transportes</i>
ANACOM	<i>Autoridade Nacional de Comunicações</i>
AOM	<i>Área Operacional Marítima</i>
APA	American Psychological Association
ASPOF	<i>Aspirante a Oficial</i>
BMM	BlueMassMed
BRITE	Baseline for rapid Iterative Transformational Experimentation
C2	Command and Control
CADOP	<i>Centro de Análise de Dados Operacionais</i>
CENTRIX	Combined Enterprise Regional Information Exchange System
CFR	<i>Capitão-de-fragata</i>
CFR	Community Fleet Register
CINAV	<i>Centro de Investigação Naval</i>
CISE	Common Information Sharing Environment
COG	Course Over Ground
COMAR	<i>Centro de Operações Marítimas</i>
COP	Common Operating Picture
CSM	<i>Conhecimento Situacional Marítimo</i>
CSV	Comma Separated Values
DGRM	<i>Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos</i>
DITIC	<i>Direção de Tecnologias de Informação e Comunicações</i>
DOTMLPII	Doctrine, organization, training, logistic, leadership, personnel, infrastructure and interoperability
DSC	Digital Selective Calling
EMC	<i>Equipamento de Monitorização Contínua</i>
EMSA	European Maritime Safety Agency
EO	Earth Observation
ETA	Estimated Time of Arrival
EU	European Union



EU CDC	European Union Cooperative Data Centre
EUROSUR	European Border Surveillance System
FASTC2AP	Fast Connectivity for Coalition Agents Program
FND	<i>Forças Nacionais Destacadas</i>
GIS	Geographic Information System
GMDSS	Global Maritime Distress Safety System
GMSK	Gaussian Minimum Shift Keying
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GUI	Graphical User Interface
I2C	Integrated System for Interoperable sensors & Information sources for Common abnormal vessel behaviour detection & Collaborative identification of threat
IMO	International Maritime Organization
ITU	International Telecommunication Union
KDD	Knowledge Discovery in Databases
KEEL	Knowledge Extraction based on Evolutionary Learning
KNIME	Konstanz Information Miner
LRIT	Long-range Identification and Tracking
MARISA	Maritime Integrated Surveillance Awareness
MARSUNO	Maritime Surveillance in the Northern Sea Basins
MATLAB	MATrix LABoratory
MCCIS	Maritime Command and Control Information System
MDA	Maritime Domain Awareness
MMSI	Maritime Mobile Service Identity
MONICAP	<i>Monitorização Contínua das Atividades da Pesca</i>
MRCC	Maritime Rescue Coordination Centre
MSA	Maritime Situational Awareness
MSSIS	Maritime Safety and Security Information System
NATO	North Atlantic Treaty Organization
NEREIDS	New Services Capabilities for Integrated and Advanced Maritime Surveillance
NGA	National Geospatial-Intelligence Agency
NIRIS	Networked Interoperable Real-Time Information Services
NM	Nautical Miles



NMEA	National Marine Electronics Association
PERSEUS	Protection of European seas and borders through the intelligent use of surveillance
PJ	<i>Polícia Judiciária</i>
RMP	Recognised Maritime Picture
ROT	Rate Of Turn
SADAP	<i>Sistema de Apoio à Decisão para a Atividade de Patrulha</i>
SAR	Search and Rescue
SAR	Synthetic Aperture Radar
SEABILLA	Sea Border Surveillance
SEF	<i>Serviço de Estrangeiros e Fronteiras</i>
SIMTISYS	Simulator for Moving Target Indicator System
SOG	Speed Over Ground
SOLAS	Safety of Life at Sea
SOTDMA	Self-Organizing Time Division Multiple Access
STI	<i>Superintendência das Tecnologias de Informação</i>
TIDE	Technology for Information, Decision and Execution
USA	United States of America
VHF	Very High Frequency
VMS	Vessel Monitoring System
VoIP	Voice over Internet Protocol
VTs	Vessel Traffic Services
WEKA	Waikato Environment for Knowledge Analysis
WGS	World Geodetic System
YALE	Yet Another Learning Environment



Data mining for anomaly detection in maritime traffic data



CHAPTER 1

INTRODUCTION

- 1.1 Motivation
- 1.2 Topic's relevance
- 1.3 Goals
- 1.4 Structure

1. Introduction

In the last decade, oceans have regained an unparalleled importance to all nations, mainly due to the financial influence they represent for each country, especially for those that are considered coastal States, with large sea borders.

Nowadays, oceans are used by almost every country as an effective and cost efficient way to commerce goods. In fact, maritime traffic has increased significantly, being considered the main route of global trade.

The United Nations refer that “the world shipping fleet provides not only transport connectivity to global trade but also livelihoods to the people working in maritime businesses in developed and developing countries” (United Nations, 2017). In fact, if the values at the beginning of 2017 are to be considered “the world fleet’s commercial value amounted to \$829 billion, with different countries benefiting from the building, owning, flagging, operation and scrapping of ships” and “with over 80 per cent of global trade by volume and more than 70 per cent of its value being carried on board ships and handled by seaports worldwide, the importance of maritime transport for trade and development cannot be overemphasized” (United Nations, 2017).

Accordingly, the growth in sea transport has consequently resulted in the intensification of traffic density. This brings new issues regarding the navigation and maritime transport at safety and security levels (Sulemane, 2015).

More specifically, Portugal, a coastal State with 92.226,0 km² (Pordata, 2016), located on the South West Europe, has a vast maritime area under its domain. As observed in Figures 1 and 2, it is easy to conclude the vast commercial routes and immeasurable amount of ships that cross the Portuguese jurisdictional waters.

According to *Autoridade da Mobilidade e dos Transportes* (AMT)¹, cargo trade at the Portuguese mainland ports grew 5.1% by the end of October 2017 to a record of 81.3 million tonnes. Lisbon stands out, with a year-on-year gain of 26% for the best record in the last nine years. By trading 10.3 million tonnes, the port of the capital alone contributed with 2.1 million tonnes to the increase in activity on the mainland (Transportes e Negócios, 2017).

¹ Transportes and Mobility Authority. It is the regulatory and oversight body for the transport sector in Portugal.

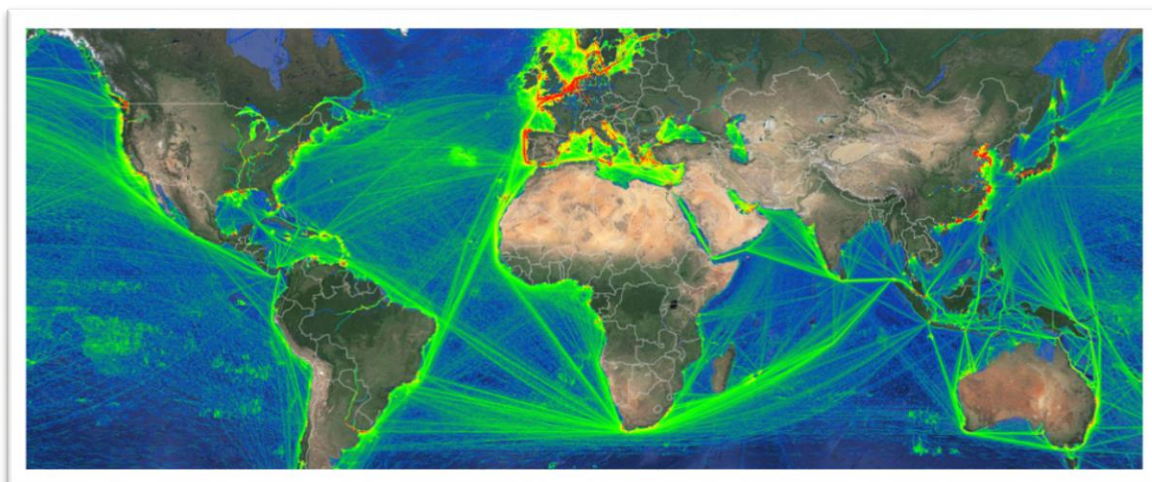


Figure 1 - Main international maritime routes (Jornal da Economia do Mar, n.d.).

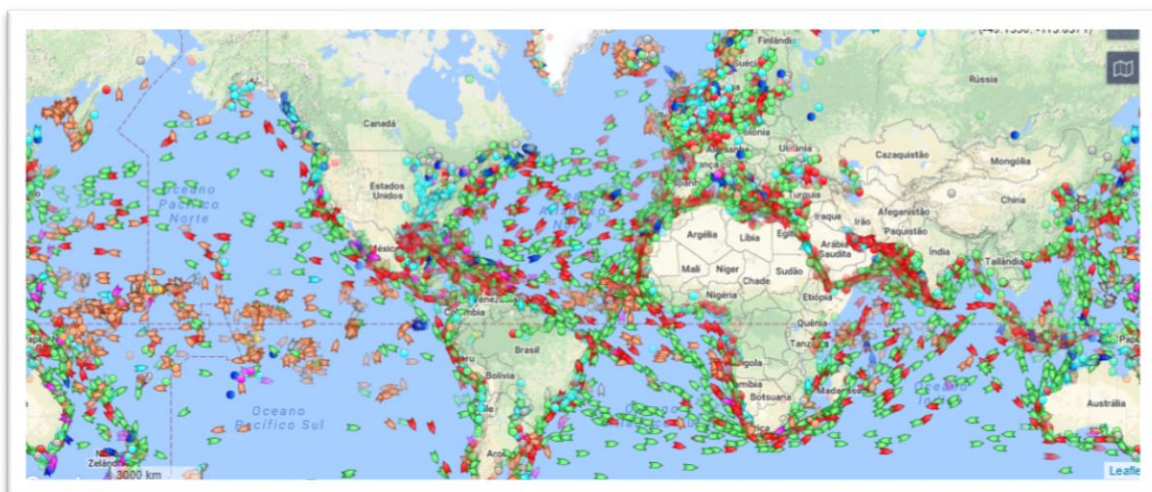


Figure 2 - Real time maritime traffic (Marine Traffic, n.d.).

In addition to trade, there are also other resources in national waters, likewise important, which are of considerable value and need to be protected and preserved. Therefore, it is clear the need to ensure the proper use of the sea, preventing its illegal exploitation by defending its interests, encouraging the overall stability. It is crucial to have an adequate monitoring of national waters, to guarantee safety and security.

Portugal is located in a privileged position, not only in terms of international maritime traffic for being near focal points of maritime traffic, but also because of its responsibility in the area of Search and Rescue (SAR). With its archipelagos located in the North Atlantic, this State is responsible for a SAR area almost 63 times bigger than

the country itself, with 5.754.848 km² (Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos, n.d.).

In Portugal, it is the Portuguese Navy that has the responsibility of ensuring that the country can use the sea for its own interests. Like other navies throughout the world, it is based on a dual concept structure. Considering this structure, on the one hand, they use their ships and means to perform in actions of military defence and in support of foreign policies. On the other hand, they use them to perform in actions concerning the authority and security of the country as well as to support the development of various areas such as scientific and cultural, thus allowing the maximization of resources. In other words, by carrying out missions of military and non-military nature.

That is why the Portuguese Navy, together with *Polícia Judiciária* (PJ)² and *Serviço de Estrangeiros e Fronteiras* (SEF)³, plays a big part in the fight against illegal maritime activities. From these illegal activities, the following stand out: terrorism (in Europe and possibly, in the near future, against the shipping lines essential for the world economy), illegal immigration and human trafficking (mostly from the North of Africa to Europe), piracy (for example near Guinea Gulf), drug trafficking (such as cocaine and haxixe), illegal fishing and depletion of natural resources, among other more (Melo, 2011).

Due to its geographical position, Portugal is in the centre of the scene of multiple illicit activities of transnational networks. Illegal activities are inextricably linked to trade by sea.

Some of these illicit activities, like drugs and weapons traffic, can be very difficult to detect. There is no specificity in the type of ship that carries them, so in order to detect them, in an appropriate amount of time, cross-agency collaboration is paramount.

On the other hand, illicit activities such as illegal fishing and unauthorised scientific research can be more easily controlled, by being specialised vessels. However, that is not an absolute truth because, depending on the size of the fishing vessels, they may or may not have to report their position, in addition to the fact that they are not as straightforwardly detected by coastal sensors.

² Judiciary Police.

³ Immigration and Borders Services.

Both the European Union (EU) and the North Atlantic Treaty Organization (NATO) are counting on Portugal means to prevent and, if need be, deter any activity that could jeopardize the safety and security of ships passing through this large area of maritime responsibility. This is not an easy task as more threats assume a diffuse feature, not easily recognisable and requiring more of the surveillance systems and their operators.

Two different perspectives can be adopted concerning threats at sea. The first, maritime safety, has as a main goal to guarantee the safety of navigation and of those who are at sea, by preventing maritime accidents and, if not possible, deal with the outcome and the consequences that may arise from those accidents. It also has the goal of protecting the maritime environment and its sustainability by preventing the exploitation of resources and pollution. The second, maritime security, aims to combat illegal actions at sea (Sousa, 2013). This kind of threats are much more complex, because they depend on third parties seeking to exploit the vulnerabilities of coastal states, in covert actions.

Taking into consideration the sea's economy growth, there is the need to create a system that has the capacity to integrate all the information coming from different sources.

The integration of all these sources of information will certainly allow to improve and increase maritime situational knowledge, contributing to a more conscious action throughout the decision cycle, making good use of time and financial resources.

By belonging to the National Emergency Service network, Maritime Rescue Coordination Centre (MRCC) allows the Portuguese Navy to give a quick and specialized response to emergency and rescue situations in its area of responsibility.

Centro de Operações Marítimas (COMAR)⁴, however, is the focal point for command and control of all the activity carried out by the Navy, having to be in constant coordination with all entities with responsibilities in marine areas under sovereignty, jurisdiction or national responsibility (Marinha, 2018).

COMAR “has an interagency and interdepartmental approach” (Marinha, 2018), not only on a national level (the Portuguese Navy, National Maritime Authority, Portuguese Air Force, SEF and PJ, among others) but also on an international level,

⁴ Maritime Operations Centre.

where several information networks coexist in order to improve the coordinated efforts of the different state departments, on their areas of jurisdiction and national responsibility. These networks include military and non-military information.

The fact that MRCC and COMAR occupy the same building is a substantial practical advantage. It allows access to different systems, therefore improving Maritime Situational Awareness (MSA), thus helping to monitor activities at sea.

When it comes to COMAR's role in the military defence functions and support of the State's foreign policy, it maintains a permanent monitoring on the missions of *Forças Nacionais Destacadas* (FND)⁵, such as those missions under the guidance of the European Union and NATO in the fight against piracy in Somalia or NATO counter-terrorism missions in the Mediterranean (Marinha, 2018).

COMAR also maintains a continuous monitoring on all Navy's vessels with assigned mission in national waters as well as on the means of the National Maritime Authority. It acts as a pillar in the management of Navy's operational information. In order to do this, it relies on the *Centro de Análise de Dados Operacionais* (CADOP)⁶.

For this dissertation, it is also important to distinguish two almost interchangeable concepts such as surveillance and monitoring.

Maritime surveillance can be understood as the systematic observation of maritime areas by all means available, with the goal of controlling the movement of vessels or other vehicles, operating on the surface or sub-surface of the seas and oceans. The ability to continuously cover large areas, with features like accuracy, data discrimination and confidence describe what surveillance is all about.

Monitoring, on the other hand, can be distinguished from surveillance considering its main goal. The aim is to maintain and improve standard and safety procedures in order to have a better understanding of the environment by using actions belonging to surveillance, with this knowledge being continuously improved over time (Carolas, 2016).

The monitoring therefore consists in controlling one or more parameters in order to detect anomalies.

⁵ National Abroad Forces.

⁶ Operational Data Analysis and Management Centre.

1.1. Motivation

Each coastal state recognises the importance of the sea resources and the importance to take all measures so that what is his remains his, defending their rights in what concerns exploitation and exclusivity. Therefore sea presence is fundamental, and due to the vastness of maritime areas under Portuguese sovereignty or jurisdiction, that may not always be an easy task, opening the opportunity for thirds parties to use resources that do not belong to them. In order to ensure that laws and regulations are enforced, coastal states should bet on improving the surveillance on their areas.

As a matter of fact, most of the activities taking place in Portugal's maritime area are not a direct, but a transitory threat. Nonetheless, it is important to avoid the absence of authority at sea, taking a special care to not disregard any surveillance measures. It is by maintaining a high level of maritime situational awareness, that Portugal can counteract any attempts of illicit activities, always being one step ahead.

1.2. Topic's relevance

The biggest problem regarding all the available data is the difficulty to transform data in information and then putting into practice all the acquired knowledge. There is so much data available that it is difficult to discern what is important and what can be put in second plan. It is when connections between data sets are analysed or context is introduced that some insight is usually gained, as it is represented in Figure 3.

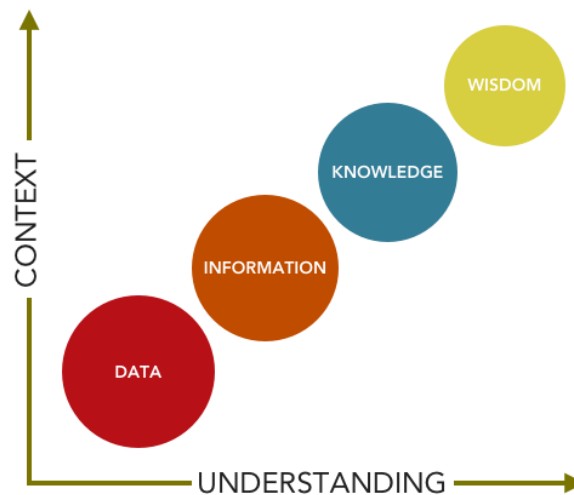


Figure 3 - Data, information, knowledge and wisdom (context vs understanding) (Marinha, 2018).

The need to integrate tools that merge data from different sources and compile them into a single system is increasingly evident. At a time when illicit trafficking in the

sea surpasses values never seen before and illegal migration is a harsh reality, there is a pressing need to introduce means to help the decision making of who is on the ground seeking guidance.

At the level of the Portuguese Navy, as an end-user, this subject gains particular prominence. Having access to a software that could respond to the diversity of challenges that the naval units come cross would be an added value.

1.3. Goals

In this dissertation, the main goal is to answer the following question:

- How to maximize the achievement of Maritime Situational Awareness by detecting anomalies using different sources of data?

However, in order to do so, it is important to gain some insight on the area and answer some questions derived from the above, such as:

- What is Maritime Situational Awareness and what is the current state of it in Portugal?
- What data sources contribute to a good Recognized Maritime Picture (RMP)? And which ones are available?
- How can the data be used to detect anomalies in the maritime traffic?

1.4. Structure

The following dissertation is divided into 6 chapters, including introduction and conclusion. Chapter 2 consists of a literature review, where concepts as Maritime Situational Awareness, anomaly and anomaly detection will be addressed as well as a brief explanation will be presented on some data mining tools. Chapter 3 will consist on a description of the methodology and all the processes the data underwent since its collection to the moment they were inserted in PostgreSQL database. Chapter 4 will describe the scope of the study area as well as all the techniques used in the data analysis. Chapter 5 will present all the tests conducted and the respective results. The conclusion, will summarize what were the results, the difficulties felt and future projects that could be taken into consideration.

In order to protect data confidentiality, all information regarding vessel's identity was modified. Any resemblance to reality is purely coincidental.

The referencing style present throughout this dissertation is American Psychological Association (APA).

CHAPTER 2

LITERATURE REVIEW

- 2.1 Maritime Situational Awareness
- 2.2 Maritime Domain Data
- 2.3 Concept of Anomaly
- 2.4 Anomaly Detection
- 2.5 Data Mining Tools

2. Literature Review

2.1. Maritime Situational Awareness

Maritime Situational Awareness is a very important concept and must be fully understood. The first thing that must be comprehended is that MSA's concept is not a static one (Pereira, 2010) and more often than not it does not result from the input of one single system but from the combination of several working together. To some authors, MSA is "the capability of understanding events, circumstances and activities within and impacting the maritime environment" (Arguedas, 2015). Other definitions go even further and say that MSA should facilitate the process of decision making and permit an effective operational response (Estado Maior da Armada, 2012).

The concept of Maritime Situational Awareness appeared based on an already existing concept, Maritime Domain Awareness (MDA), following a NATO summit in Riga, 2006. In May 2008, it was presented by NATO a MSA Concept Development Plan with the main purpose of implementing a MSA with a Doctrine, Organization, Training, Logistic, Leadership, Personnel, Infrastructures and Interoperability (DOTMLPPI) approach (Veloso, 2015).

MSA seeks to obtain, as a final result, a fully clarified surface picture, receiving to that end the contributions of several surveillance and monitoring systems as well as knowledge generated by other sources. If data is able to cover "all aspects of a situation of interest in a timely manner, one can then say that complete and continuous situational awareness has been achieved" (Martineau, 2011). However, it is important to understand that total situational awareness "would be akin to omniscience and achieving it would be a utopia" (Martineau, 2011).

In order to produce knowledge regarding the maritime domain by identification of patterns of the maritime community, reliable and continuous data must be collected from all kinds of data sources.

Therefore, the creation of safety applications that are capable of detecting behavioural patterns of ships and anomalies that indicate potential situations of infraction is not only feasible but of utmost importance. These applications might help the security forces clarifying the maritime picture. Only then will it be possible to take action in a timely manner, maximizing the use of resources and minimizing risks (Estado Maior da Armada, 2012). The goal is to have automation in the generation of alerts in order to trigger the corresponding actions.

In today's world, the prime challenge of MSA is the “aggregation of large amounts of heterogeneous data and their transformation into useful and reliable information to support users in the decision making process” (Arguedas, 2015). From the analysis of a large volume and variety of data with spatial and geographic representation, using techniques that can detect connections between events, it is possible to provide mechanisms to draw certain conclusions (Arguedas, 2015).

In fact, there are so many sources of data nowadays that the issue now is not the lack of data but the overwhelming amount of it. Therefore, there have been developed several automation processes in order to deal with that amount of data. It is even considered that “to manually pore through and to analyse the information in a bid to identify potential maritime threat is tedious, if at all possible” (Chen, 2013).

The ability to extract knowledge that is useful, but is usually hidden, from data is becoming more and more important in the 21st century (Ahlemeyer, 2014). On a computational level, the industry has suffered numerous technological advances, in both the hardware and software sectors, being able to store, process and analyse a large amount of data with an increasingly ease.

2.1.1. NATO's role

NATO developed several systems that contribute to MSA. Baseline for Rapid Iterative Transformational Experimentation (BRITE) was developed as “part of the initiative Technology for Information, Decision and Execution (TIDE) superiority”. It is a “National Geospatial-Intelligence Agency (NGA) program of record that provides a client-server system for image dissemination” (North American Job Bank International Networking, n.d.).

Maritime Safety and Security Information System (MSSIS) is another one of these systems. It is “a freely-shared, unclassified, near real-time data collection and distribution network” (Maritime Safety and Security Information System, 2008). It collects Automatic Identification System (AIS) data from ships from member countries and broadcasts them almost in real time (Veloso, 2015). Its main goal is to increase maritime security and safety by promoting a “multilateral collaboration and data-sharing among international participants” (Maritime Safety and Security Information System, 2008). This allows to have a RMP at a very low cost.

Maritime Command and Control Information System (MCCIS) also contributes to MSA, by providing as well “a high quality RMP” (Maritime Safety and Security Information

System, 2008) and giving inputs to NATO's Common Operational Picture (COP). MCCIS has appeared "as the C2 (Command and Control) tool of choice for NATO's maritime component commanders" (Germain, 1997).

Another one of these systems is Fast Connectivity for Coalition Agents Program (FastC2AP). FastC2AP is a "human-interactive, rule-based program" which focus on helping operators in specifying "vessel behaviours and characteristics that drive alerts and prompt operators to analyse those vessels further" (National Research Council, 2008).

Combined Enterprise Regional Information Exchange System (CENTRIX) is an "operational level network, supporting regional commanders and their staffs at a variety of security levels" (Mitchell, 2013) which allows the exchange of a Common Operational Picture, e-mails with attachments, web-enabled services, chat and Voice over Internet Protocol (VoIP).

Networked Interoperable Real-Time Information Services (NIRIS) is a system that "displays real time maritime, ground, air tactical and theatre missile defence data received from control reporting centres" (Kowalczyk, 2009). It contributes by transforming data into interoperable information, based on NATO and commercial standards, by offering a package of services that allow data compilation and dissemination (Veloso, 2015).

2.1.2. Europe's role

In 2008, a new project was embraced by European Union, being supported by the Council and European Parliament. Common Information Sharing Environment (CISE) was developed in order to establish itself as a system of information sharing, with the main purpose of expanding each country's maritime surveillance. It was supported by the countries' need to ensure safety and protection of the seas and oceans (Veloso, 2015).

The end result intended for CISE is to turn maritime surveillance the most accessible and coherent as possible between all individuals, not exactly to collect data. This would be accomplished by having automatic data sharing mechanisms that would allow access to relevant information in real time.

Therefore, several projects emerged in order to attain this goal.

Maritime Surveillance in the Northern Sea Basins (MARSUNO), was created with the purpose of supporting CISE, overcoming legal, technical and administrative obstacles (Veloso, 2015) in what concerns the sharing of information beyond borders. It counted with the participation of 9 countries throughout 24 months: Belgium, Finland, France, Germany, Latvia, Lithuania, Estonia, Poland and Sweden as the central partner.

BlueMassMed (BMM), on its turn, was created with the intent of “increasing the interoperability between the existing control and location systems” and “evaluate the project’s partners capability of sharing surveillance information” in the areas where they intervene. It was supported financially by the European Commission and by Portugal, Spain, Italy, France, Greece and Malt (Instituto Hidrográfico, n.d.), and intended to centre its efforts on monitoring the maritime situation in the Mediterranean and Atlantic approaches, establishing the prototype on which collaboration should be expanded at European level (Pereira, 2010).

In order to developed CISE’s security component, other 7 initiatives were financed by EU, namely, European Border Surveillance System (EUROSUR), Protection of European seas and borders through the intelligent use of surveillance (PERSEUS), Integrated System for Interoperable sensors & Information sources for Common abnormal vessel behaviour detection & Collaborative identification of threat (I2C), Sea Border Surveillance (SEABILLA), Dolphin, New Services Capabilities for Integrated and Advanced Maritime Surveillance (NEREIDS) and Simulator for Moving Target Indicator System (SIMTISYS) (Veloso, 2015).

In the Portuguese Navy, the document that regulates the definition of Maritime Situational Awareness is IOA 114 – *Conceito de Conhecimento Situacional Marítimo*⁷. It was developed to be used in support of decision making regarding maritime operations (Estado Maior da Armada, 2012). In an ever changing environment such as the sea, subject to multiple threats and security issues, being able to have the upper hand is key. In order for that to happen, it is essential to continuously monitor the maritime environment so that superiority of information (Estado Maior da Armada, 2012) is attained and can be used to protect the security and the state’s authority. This superiority can be achieved by obtaining relevant data, processing it and creating connections between the different types of data and therefore turn data into information. This superiority is only possible if there is confidence and integrity regarding the information

⁷ IOA 114 – Concept of Maritime Situational Awareness.

received. So information, on a military level, is shared on a need to know basis. Access restrictions are justifiable, however it poses as “a major problem in the domain of anomaly detection” (Martineau, 2011).

The management of the sea is not an easy feat considering how many national and international entities are involved, sharing responsibilities, dependencies and lines of authority. Each one has their own goals and concerns, using different systems to store, analyse, fuse, validate and share data (Estado Maior da Armada, 2012). Data sharing, or its lack thereof, is one of the main obstacles in the search of maritime situational awareness. It is through data that one gains information and now, 21st century, more than ever, information is a tool that grants power. That is why data sharing has usually several deterring policies. This causes big discrepancies between the several systems that help to achieve MSA. The amount and disparity of data (often not interoperable) and information generated concerning the maritime domain is often a hindrance, making it hard to coordinate and use on an operational level.

The construction of a robust MSA capability faces different challenges that need to be overcome. As it is shown on Figure 4, these challenges can be of 4 different kinds (Estado Maior da Armada, 2012).

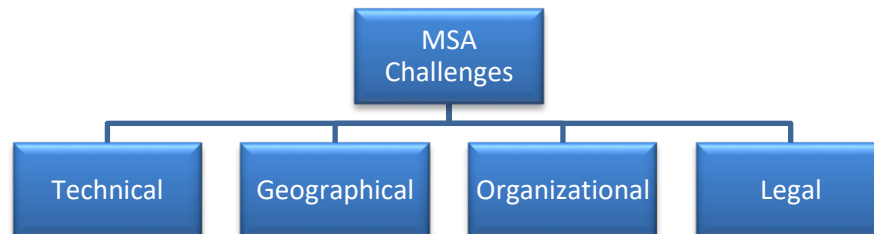


Figure 4 - Challenges of Maritime Situational Awareness

On a technical level, it is difficult to have a predefined architecture and ensure interoperability between the several systems. On a geographical point of view, the amount of sensors and systems required depend on the size and location of the area.

On an organizational standpoint, it is important to have an agile structure that allows responding to the different levels of performance in the maritime domain (Estado Maior da Armada, 2012). From a legal perspective, what immediately stands out, as



mentioned before, is the data sharing policies, because each organization or agency has its own directives, concerning data distribution, that must be followed.

Although there is often certain interferences concerning information sharing and exchange, it is extremely advantageous. Firstly, it grants access to data that would not be possible to access by own means. Secondly and lastly, because it allows to save up resources and avoid unnecessary work and effort. In this way, it is not technical matters that pose so much of a problem but the policies for data sharing.

Generating MSA involves 5 different steps: acquisition, fusion and analysis, diffusion and archive (Estado Maior da Armada, 2012).

Having an archive where all the data is stored is of particular relevance considering new algorithms and methods can be applied to that data and tested, in order to generate automatic alarms that can help the operator. This is especially important considering that “human(s) are always in the loop. Systems are made to improve the performance of operators, not to fully replace them” (Martineau, 2011).

The Portuguese Navy counts with the support of several information systems, such as Oversee. Oversee is a software developed by the company Critical Software and is considered to be a “system of systems” (Estado Maior da Armada, 2017) because it integrates information from several systems and shows it in a user-friendly manner. It was first intended to assist SAR activities, however, because it has access to classified databases and systems, such as NATO MCCIS, it can be used on an operational level, aiding to increase MSA.

It essentially receives and allows the visualization of Global Maritime Distress and Safety System (GMDSS)⁸ alerts. However, it also allows the integration of others sources of information. Systems like AIS, Satellite – Automatic Identification System (SAT-AIS), Vessel Monitoring System (VMS), also known as *Monitorização Contínua das Atividades da Pesca* (MONICAP)⁹ (European Economic Community, 1993), Long Range Identification and Tracking (LRIT) system, among others, can be inserted in Oversee.

⁸ GMDSS is an international system that through satellite technology and radio communication equipment aims to provide automated communication for ships in distress.

⁹ Fishing Activities Continuous Monitoring.

Currently, there are several projects that aim to improve MSA. The most recent one, is Maritime Integrated Surveillance Awareness (MARISA). It has, as one of the main goals, the fusion of heterogeneous data and information, creating a toolkit that allows an effective information sharing, the identification of possible risk situations and assist the process of decision making. It counts with the participation of 22 partners located in 9 European countries. Portugal assures its participation with Inovaworks (a national technological company), INESC-INOV (a development entity) and through the Portuguese Navy with *Centro de Investigação Naval* (CINAV)¹⁰. It was through CINAV that this dissertation topic was suggested in order to contribute, if possible, to MARISA project.

2.2. Maritime Domain Data

In general, data can have two origins: NATO systems or external commercial sources. Before going through several tools of fusion and analysis in order to produce information, it is necessary to compile and share data with other systems with the same degree of classification.

The data acquired on the geographic distribution of vessels can be divided into two groups, “self-reporting or observation-based” (Arguedas, 2015) according to the way the data is obtained.

There are several examples of self-reporting data, such as AIS, LRIT and VMS/MONICAP, each one of them with different purposes, such as collision avoidance, security and safety and fisheries monitoring (Arguedas, 2015).

Long-Range Identification and Tracking of ships is one of these sources and it was established in 2006 as an international system by the International Maritime Organization (IMO). LRIT system uses vessels’ satellite communications in order to transmit its information, such as position, which is transmitted at least four times a day. The European Union Cooperative Data Centre (EU CDC) tracks over 8000 ships per day and it aims to disseminate LRIT information on European ships around the world, being one of the largest LRIT distribution centres (European Maritime Safety Agency, n.d.). As a result, LRIT is one of the data sets used by the European Maritime Safety Agency (EMSA) and other organizations interfaces, being of great importance for maritime safety and awareness. AIS and MONICAP will be discussed later on, considering that, in spite

¹⁰ Portuguese Naval Research Centre.

of all these data sources, they are the ones that will be further analysed throughout this dissertation.

EMSA was developed in 2002 and it aims to decrease the risk of “maritime accidents, maritime pollution from ships and to prevent the loss of human life during the navigation” (Dorel, 2013). This is achieved with the creation of legislation, introducing measures regarding maritime safety and defence, which are generated in accordance with the already existing rules of the member states. EMSA is also responsible for providing information, upon request, to competent national authorities and European Union institutions. One example of this, is the on-going cooperation with COMAR, as EMSA provides permanent access to information regarding maritime awareness. Therefore, the collection of big sets of data is a must in order to maintain a constant monitoring of the seas.

Observation-based data, on its turn, is collected by passive or active sensors. Examples of this type of data are Synthetic Aperture Radar (SAR), space-based Earth Observation (EO) (Arguedas, 2015). These sensors’ detection capabilities depend on a variety of factors, whether they are of a more technical sort (e.g. resolution) or of an environmental nature (e.g. Sea state).

SAR main advantage is the fact that it benefits from “the long-range propagation characteristics of radar signals” (Sandia National Laboratories, n.d.) and the “information processing capability of modern digital electronics to provide high resolution imagery” (Sandia National Laboratories, n.d.). It is independent of flight altitude (providing a high resolution capability), independent of the weather (provided that the proper frequency range is selected) and it also has day/night capability due to its own illumination (Moreira, 2013).

2.2.1. AIS

AIS was developed with the main goal of providing itself “as a tool for maritime safety – vessel collision avoidance” (Tetreault, 2005), intended to be used by Vessel Traffic Services (VTS) in order to track and monitor vessel movements operating near their coasts and to assist ships’ watch-keepers. It is expected that AIS, as a VTS tool, would “be used in conjunction with traditional VTS sensors and tools” (Tetreault, 2005) to assist the VTS watch-stander to have a thorough and broad comprehension of the maritime traffic, allowing him to monitor the current situation and offer recommendations to mariners.



By providing coastal stations with the information required under “mandatory or voluntary reporting schemes as well as for VTS purposes”, AIS can help reducing “the work of the watch-keeper” (MCANET, n.d.).

Therefore, AIS is “an international standard for ship-to-ship, ship-to-shore and shore-to-ship communication of information” (Tetreault, 2005). It was regulated and normalized by the International Telecommunication Union (ITU) and implemented by IMO afterwards.

All International Requirements for carriage of AIS can be found on Chapter V, Regulation 19.2.4 of the Safety of Life at Sea (SOLAS) Convention. This regulation can be found in Annex A.

No matter the area where the ships are operating, across the ocean or near the coast, AIS works in an “autonomous and continuous mode” (Navigation Center, n.d.). For transmission, only one radio channel is necessary. The radios associated with the AIS consist of “1 full-range Very High Frequency (VHF) transmitter, 1 Digital Selective Calling (DSC) Ch. 70 receiver (used for frequency management and DSC polling), and 2 VHF Gaussian Minimum Shift Keying (GMSK) receivers” (Tetreault, 2005). However, in order to prevent problems regarding interferences in communication, each station transmits, as well as receives, more than 2 radio channels, with the following frequencies: “161.975MHz (AIS1) and 162.02MHz (AIS2)” (Melo, 2011) with the added possibility of shifting channels without any loss of communication.

There are two types of AIS, class A and class B. Class B AIS was designed as “a more economical, of smaller reach and of limited information transmitted alternative” (Melo, 2011). When compared to Class A AIS, it has an inferior reporting frequency.

AIS diverges from other kind of maritime equipment because it uses a protocol called Self-Organizing Time Division Multiple Access (SOTDMA). By not having to rely upon receiving any kind of stimulus for broadcasting, it can be set apart from other kind of equipment.

“AIS equipment self-organizes its broadcasts” (Tetreault, 2015) so that there is no interference regarding messages between AIS equipment operating in close proximity.

Each AIS station establishes “its own transmission schedule (slot)” taking into consideration “data link traffic history and knowledge of future actions by other stations” (Navigation Center, n.d.).

Every 60 seconds, 2250 time slots are established. As there is two frequencies, overall there are “4500 slots per minute available” (Melo, 2011). A simple position report regarding one AIS station is included into one of those 2250 slots. Each AIS unit automatically “determines what slots are available for its use, broadcasts its intentions for slot use to other units to allocate the slots, and transmits its messages” (Tetreault, 2015). Figure 5 represents this process.

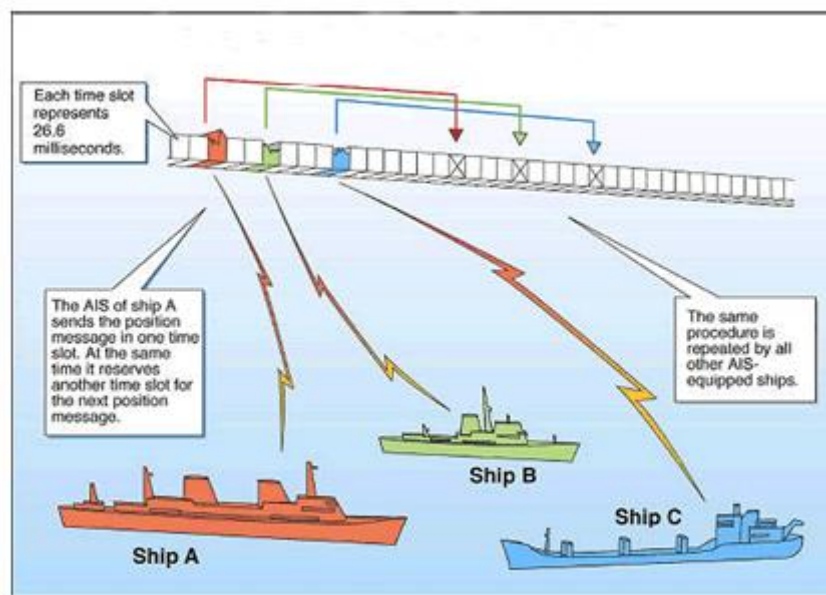


Figure 5 –AIS Functioning – Slots (Navigation Center, n.d.).

Overlapping of slot transmissions is easily avoided by constant synchronization. The selection of each slot is “randomized within a defined interval, and tagged with a random timeout of between 0 and 8 frames” (Tetreault, 2015). Before changing its slot assignment, each station announces the new location and the respective timeout.

Regarding coverage, it depends on the height of the antenna, being AIS coverage similar to other VHF applications. Compared to radar, AIS has a slightly better performance when it comes to propagation. It is not usually affected by the weather, only by the “shielding of the transmitted signal by land masses and buildings” (International Maritime Organization, n.d.). In spite of this, its wavelength is significantly longer than radar, being able to reach around the contour of the land and even in certain circumstances behind small islands, if they are not too high.



In general, at sea, AIS has a range of 15 to 20 nautical miles. Base stations, on its turn, can expand the range up to 40 to 60 miles (Marine Traffic, n.d.), provided that aspects as the weather conditions, elevation, antenna type and obstacles surrounding the antenna are favourable.

AIS can broadcast three types of information (Figure 6):

- Static, which includes the ship's name, Maritime Mobile Service Identity (MMSI)¹¹, ship type, ship size, among others;
- Dynamic, such as the ship's location (latitude and longitude), its Rate of Turn (ROT), Speed Over Ground (SOG), Course Over Ground (COG);
- Voyage related data such as vessel destination, cargo nature and Estimated Time of Arrival (ETA).

¹¹ MMSI is a series of nine digits and its purpose is to uniquely identify stations, whether they are ship stations or coast stations. To each ship corresponds a different MMSI.

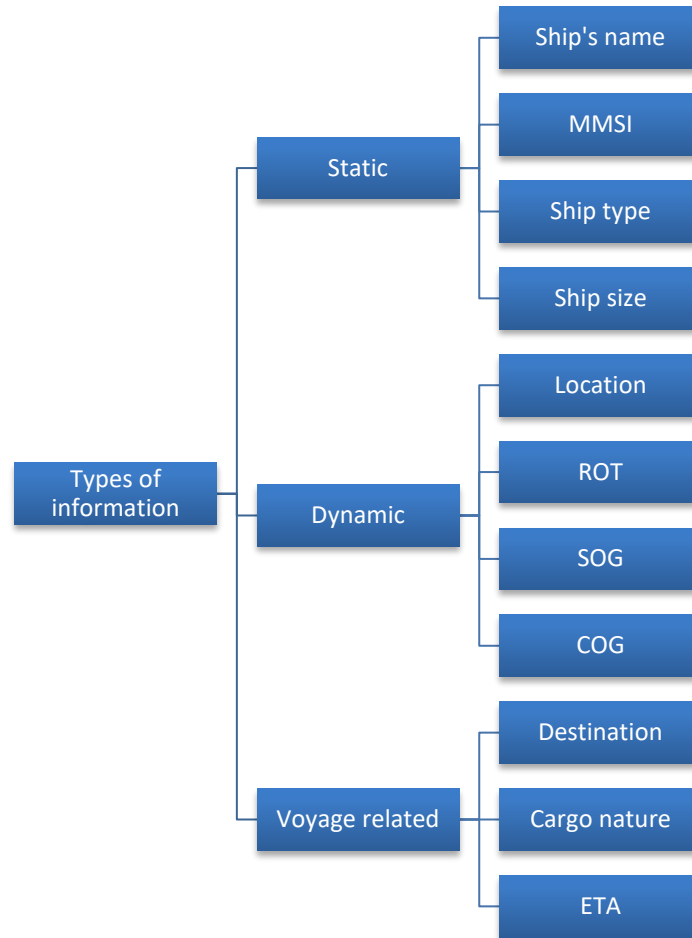


Figure 6 – Example of information given by AIS.

These types of information are compiled into messages that are autonomously broadcast “at a regular time interval” and “can be received by on board transceiver and terrestrial and/or satellite base station” (Mao, 2016). Static information, as well as voyage related information is broadcast “every 6 minutes” (Tetreault, 2015) or as requested. On its turn, dynamic information is transmitted almost instantaneously, “every 2 to 10 seconds” increasing to “every 3 minutes” if the vessel is anchored, depending on each AIS class.

There is, however, some technical issues that arise when it comes to deal with AIS data.

Firstly, it is of utmost importance that all the information contained therein must be correctly inserted (concerning static and voyage related information) and functioning properly (when it comes to dynamic inputs).

Secondly, taking into consideration that there is a vast amount of data available, handling it is no easy feat. Making it useful for obtaining Maritime Situational Awareness is one of the main goals, but for that to happen several steps must be taken. Data validation, data correlation and data fusion are some of those steps. The integration with different systems is favourable but is not as straightforward as it may seem. Moreover, the storage of such a big volume of data and making it available for being analysed by expertise is another challenge that must be faced.

There are many AIS providers. One of them is Marine cadastre, which has data collected by the United States of America (USA) Coast Guard. It contains data from 2009 to 2014 on United States' waters. Another one is Sailwx, which contains data related to a small portion of the ships worldwide. Only "those that participate in the World Meteorological Organization's program of voluntary at-sea weather reporting, and those vessels operating AIS transponders within range of a participating shore station" (Sailwx, n.d.) are contained in Sailwx database. Another one is Marine Traffic. This last one differs from the others by providing historical data at a certain cost, in relation to the volume of data requested and the processing desired (Marine Traffic, n.d.).

The data is delivered in Comma Separated Values (CSV) format and can be from terrestrial AIS receivers or Satellite AIS Receivers. As the name indicates, CSV files use the comma as a separator for each value in a text file. It can also use tab or semicolon to separate values. CSV files have several advantages, such as being a simple file, opened by text editors and easily manipulated in a programming level. It also consumes less memory than Excel files.

Although there are many AIS data providers, there is no "standard AIS benchmark database in maritime research area" (Mao, 2016). This is a great inconvenience considering all the time, effort and money involved to have a usable dataset which can be used later on by different researchers. It also gains importance considering that there is no way of comparing different methods or algorithms for anomaly detection or motion prediction because there is not a unique database to test them.

"The Portuguese Navy has a network of AIS stations" (Soares, 2012). The data obtained from those stations is then directed "to a server in the Navy's private network designated AIS server" (Soares, 2012).

The AIS server is located in Lisbon, more precisely in *Direção de Tecnologias de Informação e Comunicações* (DITIC)¹², that is an organ of *Superintendência das Tecnologias de Informação* (STI)¹³ that assures the exercise of the technical authority in the field of Navy's communication and information systems (*Chefe do Estado Maior da Armada*, 2016). It receives, through Portuguese Navy communications' infrastructures, AIS messages that compiled enable the assessment of the world panorama.

This raw data, which comes in National Marine Electronics Association (NMEA) format, is collected in MATLAB (MATrix LABoratory) files, of the type *.mat, containing what corresponds to 10 minutes of raw AIS data and is, later on, compiled in a daily file, acknowledged as "F" type file (Melo, 2011), F of footprint. Through *Sistema de Apoio à Decisão para a Atividade de Patrulha* (SADAP)¹⁴ it is possible to have access to the files compiled in a daily basis. This system will be further analysed in the next section of this document.

As it is represented in Figure 7, each daily file is divided in two sub-matrixes, one that contains reports inside the Maritime Operational Area¹⁵ (footprint_aom.mat) and other that includes reports outside of it (footprint_n_aom.mat) (Melo, 2011).

¹² DITIC - Directorate of Information and Communication Technologies.

¹³ Superintendence of Information Technologies.

¹⁴ SADAP – Activity Patrol Decision Support System.

¹⁵ Área Operacional Marítima (AOM).

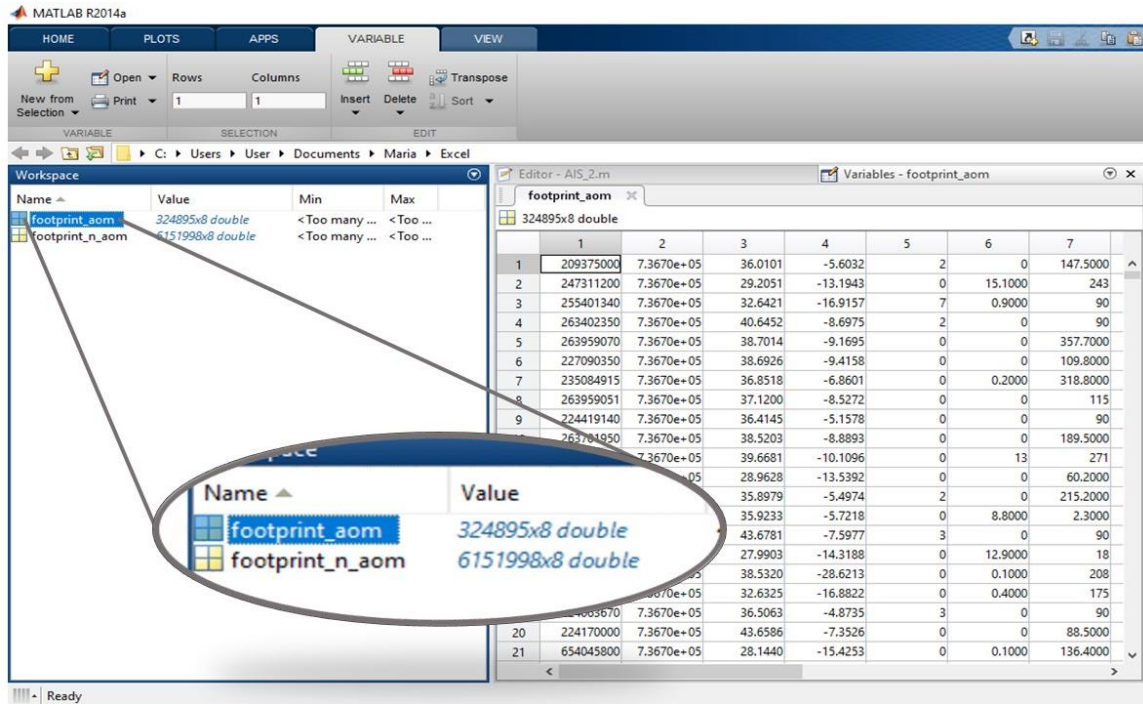


Figure 7 - AIS data, reports inside the Maritime Operational Area.

2.2.2. MONICAP

Another data type used in this dissertation is MONICAP. MONICAP or VMS, made its first appearance in 1987 (Bhargava, 2012) and is a system which aims to monitor the inspection of fishing activities (INOV, n.d.).

As it is shown in Figure 8, it uses “Global Positioning System (GPS) for the location and Inmarsat C for satellite communications between vessels and a ground control station” (INOV, n.d.).

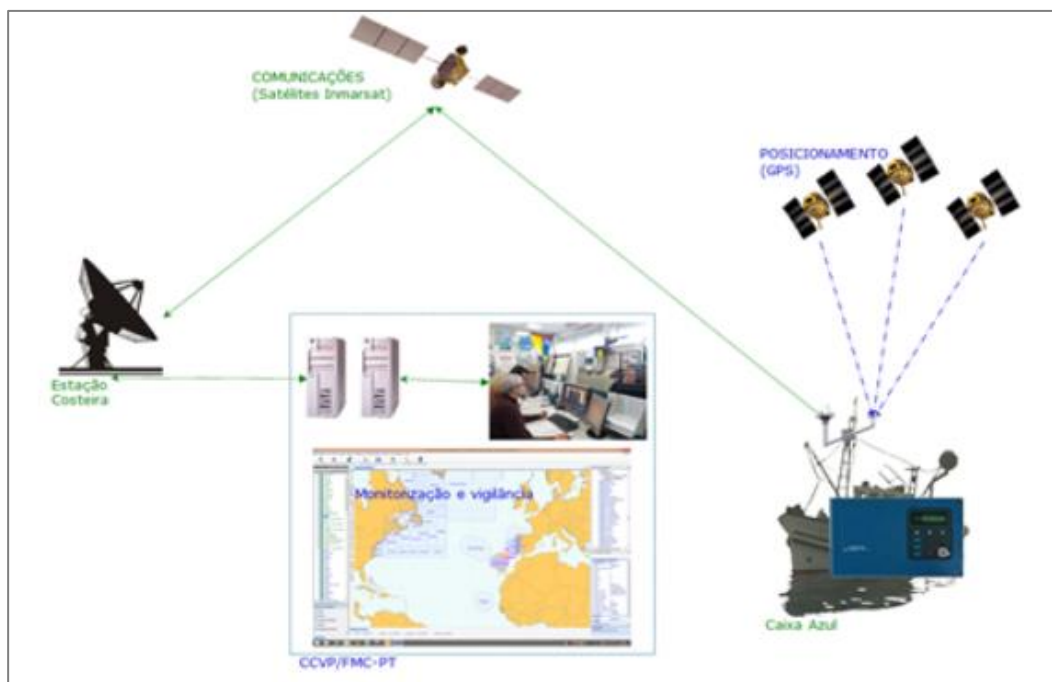


Figure 8 - MONICAP system's functioning (Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos, n.d.).

MONICAP is currently “in operation in six countries, monitoring daily thousands of vessels” (XSealence, n.d.). According to Portuguese Law Decree 310/98, article 3, line b), all ships with an over-all length of 15 meters are required to have this system on board, regardless of the place where the vessels are carrying out their activities and the fishing gear they possess.

It is based on “telecommunications technologies and geographic information” and it consists of a “continuous monitoring equipment¹⁶ installed on fishing vessels, also labelled (...) as blue box” (European Economic Community, 1993). Figure 9 represents the designated “blue box”.

¹⁶ Equipamento de Monitorização Contínua (EMC).

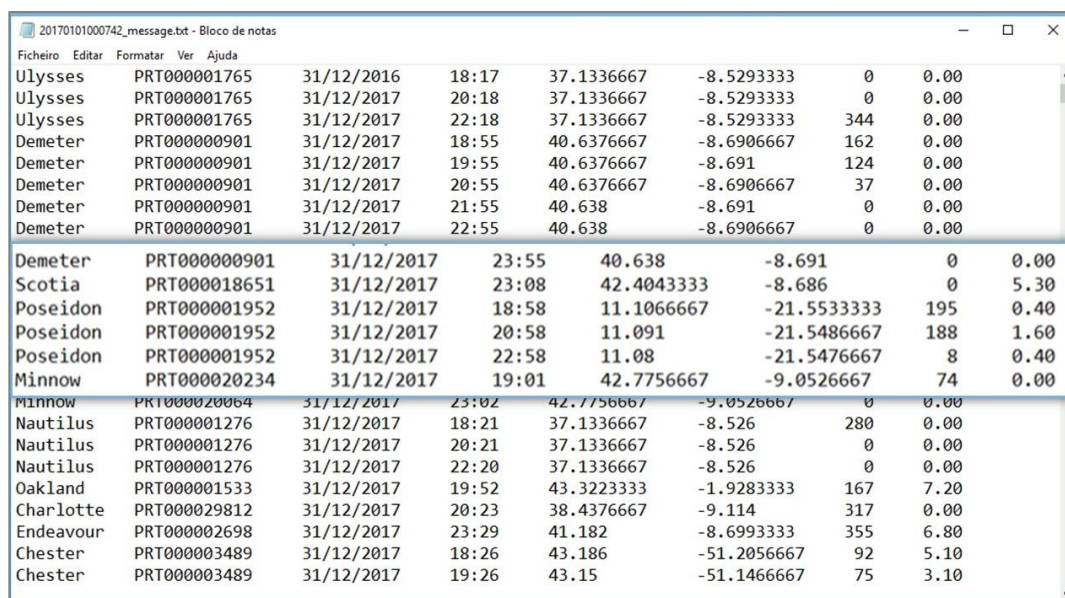


Figure 9 - Continuous monitoring equipment (XSEALENCE, n.d.).

It allows monitoring from shore, position and speed of ships on which the MONICAP box is installed. As it is shown in Table 1 and Figure 10, the MONICAP system sends data in the following format:

Table 1 – Example of MONICAP data format.

Vessel name	Community Fleet Register (CFR) number	dd-mm-yy	Hh:mm:ss	Longitude	Latitude	COG	SOG
Altair	PRT 000022323	24-01-2017	02:51:00	-27.32461232	62.14296321	089	7.40



Ulysses	PRT000001765	31/12/2016	18:17	37.1336667	-8.5293333	0	0.00
Ulysses	PRT000001765	31/12/2017	20:18	37.1336667	-8.5293333	0	0.00
Ulysses	PRT000001765	31/12/2017	22:18	37.1336667	-8.5293333	344	0.00
Demeter	PRT000000901	31/12/2017	18:55	40.6376667	-8.6906667	162	0.00
Demeter	PRT000000901	31/12/2017	19:55	40.6376667	-8.691	124	0.00
Demeter	PRT000000901	31/12/2017	20:55	40.6376667	-8.6906667	37	0.00
Demeter	PRT000000901	31/12/2017	21:55	40.638	-8.691	0	0.00
Demeter	PRT000000901	31/12/2017	22:55	40.638	-8.6906667	0	0.00
Demeter	PRT000000901	31/12/2017	23:55	40.638	-8.691	0	0.00
Scotia	PRT000018651	31/12/2017	23:08	42.4043333	-8.686	0	5.30
Poseidon	PRT000001952	31/12/2017	18:58	11.1066667	-21.5533333	195	0.40
Poseidon	PRT000001952	31/12/2017	20:58	11.091	-21.5486667	188	1.60
Poseidon	PRT000001952	31/12/2017	22:58	11.08	-21.5476667	8	0.40
Minnow	PRT000020234	31/12/2017	19:01	42.7756667	-9.0526667	74	0.00
Minnow	PRT000020064	31/12/2017	23:02	42.7756667	-9.0526667	0	0.00
Nautilus	PRT000001276	31/12/2017	18:21	37.1336667	-8.526	280	0.00
Nautilus	PRT000001276	31/12/2017	20:21	37.1336667	-8.526	0	0.00
Nautilus	PRT000001276	31/12/2017	22:20	37.1336667	-8.526	0	0.00
Oakland	PRT000001533	31/12/2017	19:52	43.3223333	-1.9283333	167	7.20
Charlotte	PRT000029812	31/12/2017	20:23	38.4376667	-9.114	317	0.00
Endeavour	PRT000002698	31/12/2017	23:29	41.182	-8.6993333	355	6.80
Chester	PRT000003489	31/12/2017	18:26	43.186	-51.2056667	92	5.10
Chester	PRT000003489	31/12/2017	19:26	43.15	-51.1466667	75	3.10

Figure 10 - Example of MONICAP messages.

The “blue box” sends this data “through satellite to a coordinator centre in Norway” (Soares, 2012). The data is then distributed to the correspondent country, taking

into consideration the country's maritime jurisdiction when it comes to fishing activities (Soares, 2012). In Portugal, the organ responsible for receiving the data is the Ministry of the Sea, through *Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos* (DGRM)¹⁷ which in turn sends it to the Portuguese Navy, more specifically DITIC. MONICAP data is stored in SADAP information system, which was developed by the Portuguese Navy in 2006 (Melo, 2011). SADAP also allows access to "statistics regarding fishery inspection activities and access to the latest position of fishing vessels operating in the Portuguese Exclusive Economic Zone (EEZ)" (Soares, 2012). Every 2 hours, the position of all vessels equipped with MONICAP is sent to DITIC through e-mail.

Besides allowing the monitoring of the fishing activities, it allows the reception and transmission of meteorological data and assists in the control of maritime traffic.

MONICAP's biggest advantage is its high reliability. It has several mechanisms that guarantee that the box equipment is not violated by presumable transgressors who would want to cover up illicit activities. That is what separates MONICAP from AIS data, the permeability to external interference. One disadvantage is the way the data is received, via e-mail, streaming near real-time would be best. In addition, the field with the vessel name can be vague considering that the names can repeat themselves. However, with the appearance of the Community Fleet Register number (CFR) that no longer poses a problem.

AIS and VMS are disparate and independent "tracking systems". Although each one of them has its own advantages, by correlating both data sources it is possible to accurately build a specific track of a ship and identify its activity. AIS can be used to deal with the VMS outages, which is not a rare event, and VMS can help corroborating some of the AIS data. The main differences between AIS and VMS are presented in Table 2.

¹⁷ General Directorate for Natural Resources, Safety and Maritime Services. It results from the fusion of *Instituto Portuário e dos Transportes Marítimos* (IPTM) with *Direção-Geral das Pescas e Aquicultura* (DGPA).

Table 2 – AIS and VMS comparison (Navigation Center, n.d.).

	Automatic Identification System (AIS)	Vessel Monitoring System (VMS)
System Type:	Digital VHF-based radio system	Satellite-based
Service Provider:	Open, non-proprietary protocol	Closed, proprietary protocols
Range:	2-way exchange of info between ships and ship-shore	Primarily 1-way (ship-shore) either scheduled or manual
Use:	Line of Sight (20-40 nm)	Line of Sight (with satellite, not ground station)
Applicability:	REQUIRED per SOLAS V/19.2.4 or 33 CFR 164.46 (NLT 2005 on certain vessels)	REQUIRED on some fishing vessels (~2000)

2.3. Concept of Anomaly

Firstly, one must clarify that there is no consensus on the meaning of anomaly. From the dictionary, anomaly can be considered a “deviation from the common rule, type, arrangement, or form” or even as “an incongruity or inconsistency” being often a synonym of “peculiarity” (Dictionary, 2018). The meaning of pattern is also relevant here. By definition, “a pattern is composed of recurring events that repeat in a predictable manner” (Martineau, 2011). What can be predicted is often considered as “normal” and what cannot, is considered as anomalous.

However, the boundary between what is regarded as normal and abnormal is very thin, considering that the prediction of normal behaviours is not an easy task. This is especially true when the “anomalies are the result of malicious actions” (Chandola, 2009), with the goal of masking anomalous observations as normal. It should be kept in mind that what is normal is often changing and evolving.

Two concepts that often are used undistinguishably is abnormality and threat. Although usually threats are anomalies, not all anomalies are considered threats. In fact, for some, “an anomaly is only a threat in context” (Seibert, 2009). To be able to classify an anomaly, whether or not it poses a threat, is an advantage because it allows to handle them in order of priority.

Several types of anomalies can be considered, falling into the following categories:

- Point anomalies: if a certain data instance stands out from the rest of the data as being anomalous then it is considered a point anomaly.
- Contextual Anomalies (or conditional anomalies): data instances that under a certain context gain another dimension and fall out of the ordinary. It is only when

the context is introduced that the data may be classified as anomalous, otherwise it does not stand out.

- **Collective anomalies:** in this type of anomaly an individual data instance on its own does not attract attention as an anomaly. However, the occurrence of a group of data instances together may be considered as anomalous. Figure 11 helps to better understand this type of anomaly. It shows the output of a human electrocardiogram and the highlighted region represents the anomaly: a low value for an abnormally long time (Chandola, 2009).

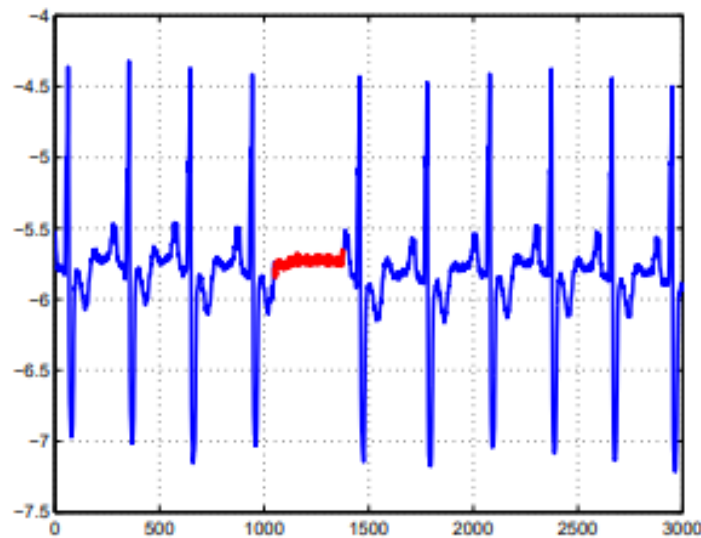


Figure 11 - Collective Anomaly example (Chandola, 2009).

2.4. Anomaly Detection

Although the majority of the systems were developed with the purpose of raising situational awareness, they can be used in anomaly detection. Due to the amount of data being generated every day, it is necessary to “sift through large quantities of data and highlight elements worthy of interest” (Martineau, 2011).

The problem of anomaly detection can be quantified as a “two class classification problem” (Gonzalez, 2002), as normal or abnormal.

One of the things to be done first is to identify extreme values in the dataset or data inconsistencies. An example of this can be an AIS transmitter that stops its transmission in an area that supposedly has coverage or AIS records that register abnormally high speeds.

Data fusion is considered to be the second step in the anomaly detection process (Martineau, 2011). This can be of especial importance given the fact that data is not always reliable and can be incomplete. Ships' position is "a classic example of fusion in the maritime domain" (Martineau, 2011). There is a multitude of ways of tracking a ship, through several sensors. Some imprecise, others that can be forged. The goal is to achieve a position as close as possible to reality. When trying to correlate two or more data sources, an anomaly can be found if there are big discrepancies between them. However, like mentioned before, it is always important to put anomalies into context, because sometimes what appears to fall out of the pattern of normality can be easily justified, for example by the environmental conditions. A practical example of correlation with different sources of data can be observed when NATO ships are conducting military operations, such as Operation Sea Guardian. This operation aims to support the maritime situational awareness and contribute to maritime security capacity-building, among other tasks.

2.5. Data mining tools

With the advent of technology, the amount of data stored has continuously increased and this tremendous rate of increment tends only to keep growing every day. It is said that human knowledge has an exponential curve, doubling roughly every century until 1900's and having a massive growth so far, with expectations that it will double every 12 hours in the future (Schilling, 2013). In this manner, there is an urgent need to have mechanisms and tools to assist the human operator extracting the useful information from such volumes of data (Fayyad, 1996). It is important to have a suitable answer. And that answer lies with data mining.

First of all it is necessary to define what is data mining and its main goals and its applicability.

Data mining is considered to be "the search for new, valuable, and nontrivial information in large volumes of data" (Kantardzic, 2011) or, in other words, data mining is "the process of discovering interesting patterns and knowledge from large amounts of data" (Han, 2012). While some people consider data mining to be the same as "Knowledge Discovery in Databases (KDD)" (Oracle, n.d.), others think that Data Mining is just a "step in the process of knowledge discovery" (Han, 2012).

This process can be divided into 7 steps (Han, 2012) which are identified in Figure 12.

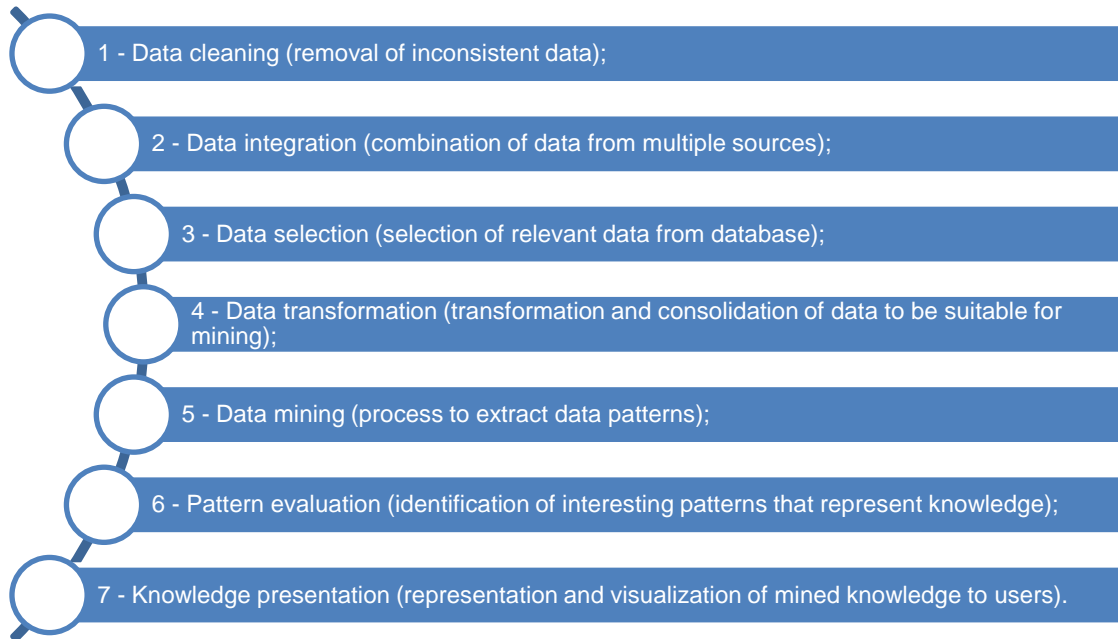


Figure 12 - Knowledge discovery process.

The tasks of data mining can be considered twofold (Shah, 2017). One of them is to create “a predictive power using features to predict unknown future values of the same or other feature” and the other one is to “create a descriptive power” (Shah, 2017) able to find a pattern that describes the data and can be easily perceived by users.

Data mining uses several different techniques to handle the data. Some of the most useful are classification and regression (both predictive tasks) and clustering and association rule discovery (of the descriptive kind).

Nowadays, there are numerous data mining tools which have as a main goal to ease the work of any person wanting to analyse big sets of data. Through tools that allow to easily gain useful insight of the data and making projections, it is possible to enhance the efficiency of the work at hands. These tools have another advantage, it is not necessary to implement standard algorithms from scratch, which allows the opportunity to test different techniques in a short amount of time. Moreover, if need be, it is also possible to change the code of the tool to fit the user’s requirements.

Presently, RapidMiner, Waikato Environment for Knowledge Analysis (Weka), R, Scikit-learn, Konstanz Information Miner (KNIME), Orange, Knowledge Extraction based

on Evolutionary Learning (KEEL) and Tanagra are some of the existing data mining tools used all over the world.

Three of them were explored and will be discussed further on.

2.5.1. RapidMiner

Formerly known as YALE (Yet Another Learning Environment), RapidMiner was first developed in 2001 in Java. This tool incorporates several data mining functions, such as data pre-processing, visualization, predictive analysis, clustering, among others. RapidMiner environment can be observed in Figure 13. All those functionalities combined with the advantage of being easily integrated with other data mining tools, like WEKA, made it a candidate for further exploring. However, the free version, the RapidMiner Studio Free Edition, has the downside of being limited to 1 logical processor and 10,000 data rows, which presents itself as a constraint when working with a great amount of data.



Figure 13 - Rapid Miner environment.

2.5.2. WEKA

Waikato Environment for Knowledge Analysis was first released in 1997 and was written in Java. As main features, it has a comprehensive set of pre-processing tools, learning algorithms and evaluation methods, different graphical user interfaces and an environment for the comparison of learning algorithms (Rehman, 2009). WEKA environment can be observed in Figure 14.

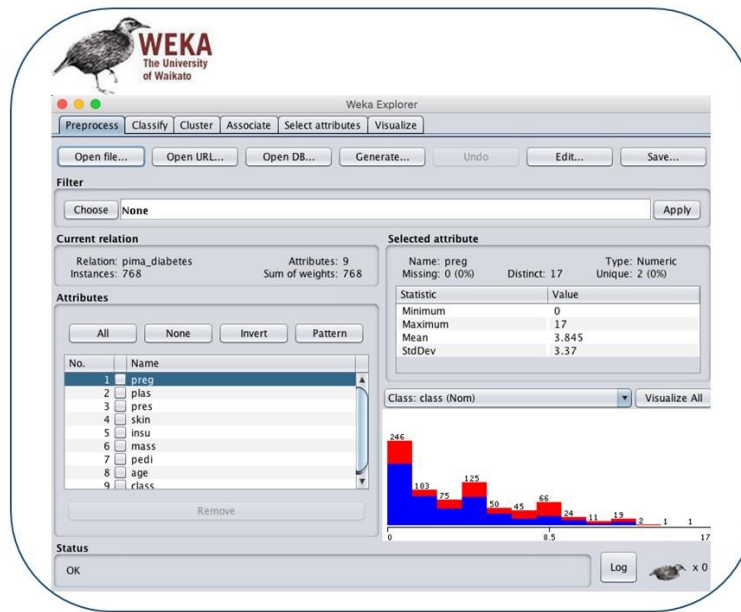


Figure 14 - WEKA environment.

One particularity of Weka is that it only deals with “flat” files¹⁸.

It also allows to choose and launch a specific Weka environment, through the Weka GUI (Graphical User Interface) Chooser as it can be seen in Figure 15.

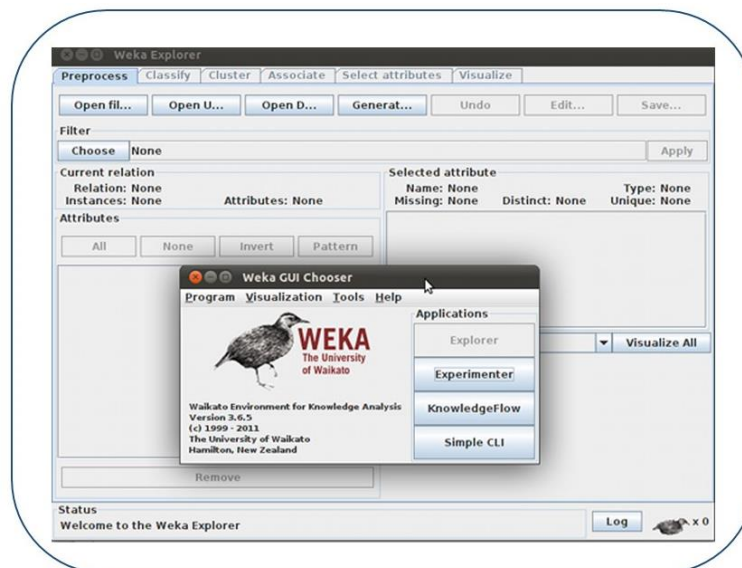


Figure 15 - Weka GUI Chooser.

¹⁸ Flat files are characterised for being simple data files in text or binary format, from which all word processing or other structure characters have been removed. CSV is an example of a flat file.

For being a tool used in classes in the Portuguese Naval Academy, it was important to further explore its capabilities.

2.5.3. Orange

Orange data mining tool has been under development since 1996 at the Bioinformatics Laboratory at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. It has a canvas interface onto which it is possible to “place widgets and create a data analysis workflow” (Bennett, 2018). Figure 16 represents Orange environment.

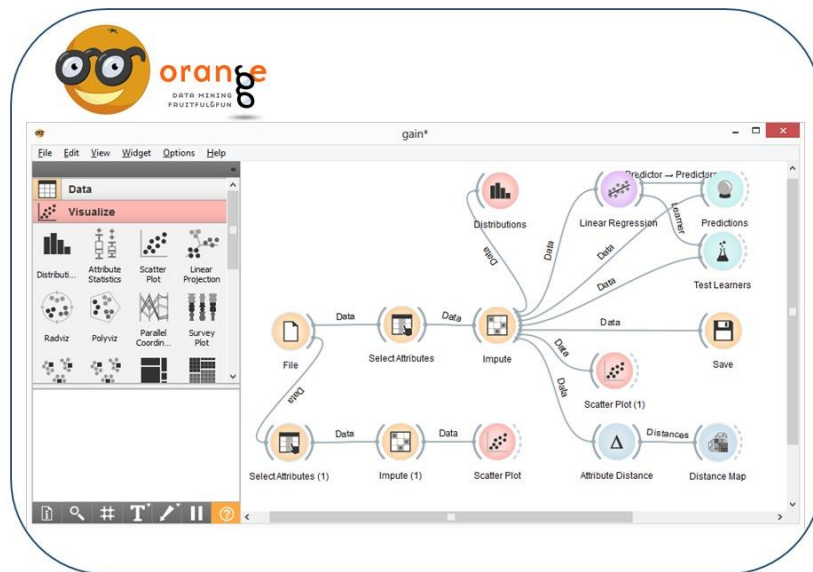


Figure 16 - Orange environment.

Widgets are considered as building blocks to create workflows and allow the user to do several functions like “reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements” (Bennett, 2018), among others. Data mining can be done in two different ways, through visual programming (using the drag and drop widgets) and through Python scripting. Widgets are divided into 5 main groups: Data, Visualize, Model, Evaluate and Unsupervised, being possible to add more groups through add-ons installation (Figure 17). The continuous development of this tool, combined with its user-friendly interface, highlights the need for further exploration.

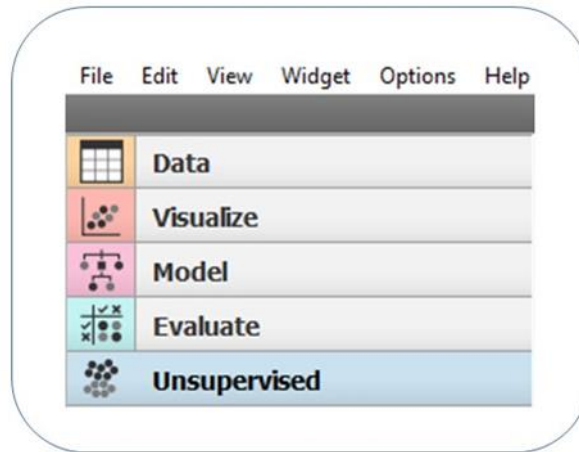


Figure 17 - Orange main widget groups

There are several studies that use one and sometimes more than one data mining tool to help to achieve a simple solution to the problem at hands.

In fact, there can be found work of researchers that use these tools to try to achieve the same results as those resulting from field experts' observations. With Orange, it was possible to build a predictive model that tries to estimate the probability that one tumor is organ confined (Zupan, 2001). This was applied to a problem of Urology and it mainly helped on the decision making by indicating the level of curability of the prostate cancer and which surgical techniques were the most appropriate.

Also through Orange, it was possible to develop an automated system which aims to identify potentially illicit elephant ivory items being sold through the multinational e-commerce corporation eBay (Hernandez-Castro, 2015).

Table 3 summarises and presents additional information about the three data mining tools explored.

Table 3 - Summary table of the 3 data mining tools explored (Predictive Analytics Today, n.d.).

Tools	Characteristics	Programming language	Operating system	Price/license
RapidMiner	Provides functionalities to optimize data exploration. Has an environment for data preparation, machine learning, among others.	Java Python	Windows macOS Linux	Proprietary software
WEKA	Collection of machine learning algorithms for data mining tasks. Includes pre-processing, classification, regression, among others.	Java	Windows macOS Linux	Open Source software
Orange	Allows visualization and analysis of data. Data mining is done through visual programming or python scripting. Has components for machine learning, text mining, among others.	Software core: C++ Extensions and query language: Python	Windows macOS Linux	Open Source software



CHAPTER 3

WORKFLOW AND DATA PROCESSING

- 3.1. Workflow
- 3.2. Data collection
- 3.3. Data Processing
- 3.4. PostgreSQL Database

3. Workflow and data processing

3.1. Workflow

Before starting the practical sections of this dissertation, it is important to delineate the steps that will be taken. Therefore, Figure 18 represents the steps of this research workflow.

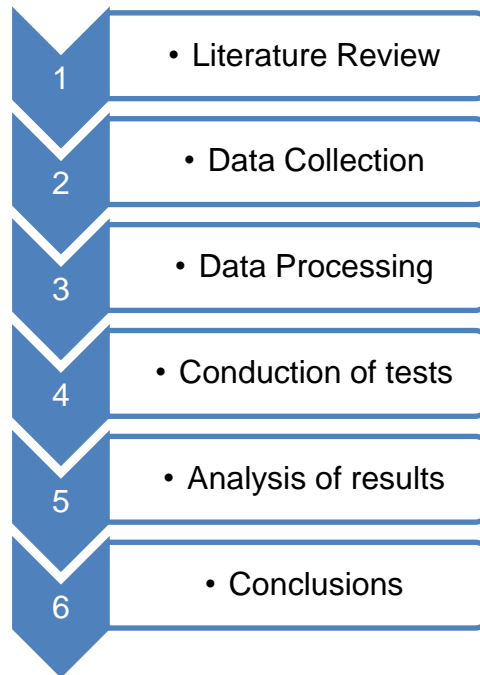


Figure 18 - Steps of the research.

Literature review was the first stage of this investigation, and it was necessary to introduce concepts that were fundamental in order to fully understand the following work.

Data Collection consists in the process of gathering AIS and MONICAP data in their source format, going through all the bureaucratic issues related to data sharing.

Data Processing will involve all the steps needed to convert data from its source format to the desired table format of PostgreSQL.

The fourth stage of the process consists in the conduction of tests. This is a fundamental part of the workflow, where correlation between AIS and MONICAP will be tested as well as the detection of specific anomalies through the chosen data mining tool.

The analysis of results is the fifth stage. The main goal of this step is to display the results obtained in the previous tests and analyze them in order to verify their integrity.

On the last step, conclusions will be drawn from the analysis of results as well as from all previous work.

3.2. Data collection

“Data is the raw material of anomaly detection” (Martineau, 2011). Without data or of poor quality, all the work done afterwards can be compromised. This makes data collection the first step and one of the most important.

As mentioned before, vessel traffic data, such as AIS and MONICAP, can be attained through Portuguese assets.

Therefore, it is possible to have access to the AIS data in .mat format through Navy’s intranet link: <ftp://ais-compiler/data>.

The MONICAP data, in .txt format, was also accessed through a Navy’s intranet link: \\PRD-MAP-APP1\CsmDataShare.

The data collected concerns the period from January 2017 to December 2017, which corresponds to the most recent available data.

Of special relevance is the fact that these sources of data have different transmission rates, with AIS data being clearly in advantage when compared with the 2 hours’ rate of MONICAP data.

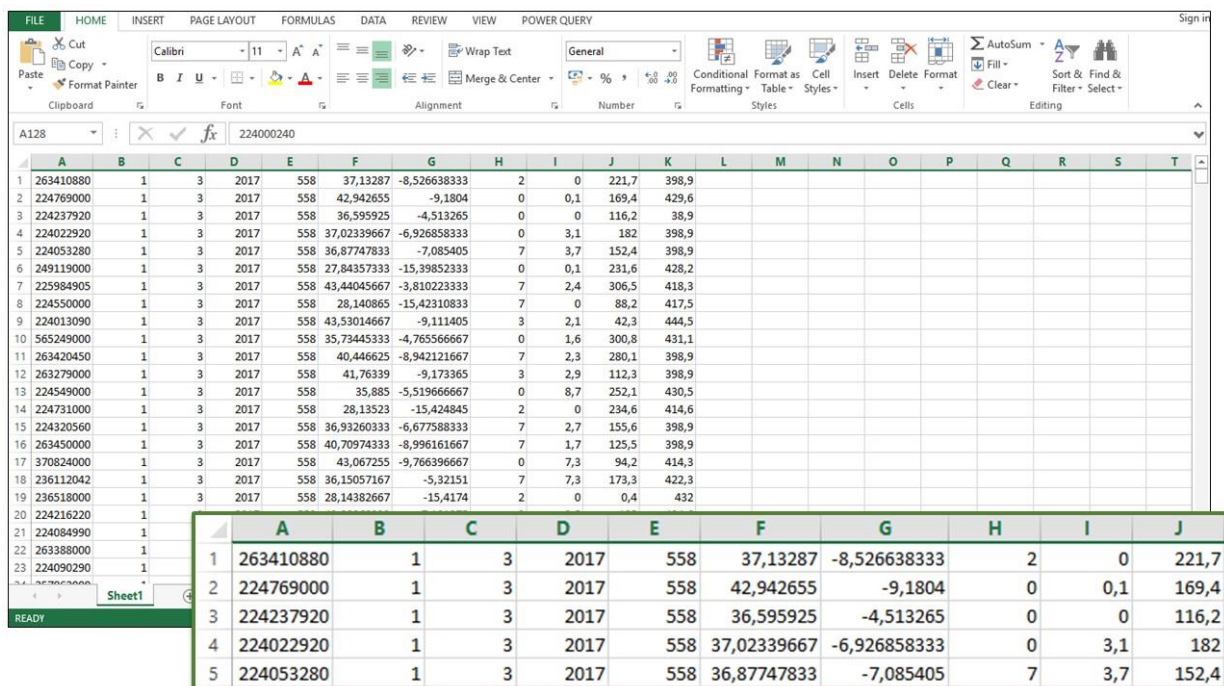
3.3. Data Processing: AIS and MONICAP

It is necessary to have a swift access to historical data and have it in a format that is easy to use in order to take full advantage of it. The main goal is to have all the data stored in the same format. PostgreSQL database was chosen to store all the AIS and MONICAP data. PostgreSQL is an open source object-relation database which is able to run on all major operating systems (PostgreSQL, n.d.).

Through a spatial database extension for PostgreSQL, PostGIS, it is possible to run spatial queries in SQL, allowing additional data types: geography and geometry (PostGIS, n.d.).

In order to store the data in PostgreSQL, it is necessary to complete a few steps before. Like it was mentioned above, AIS data comes in .mat format divided in daily files. The first step consisted in creating a simple MATLAB code (Appendix A) to have the data

files in .xlsx format¹⁹. It was necessary to tackle the fact that the date and time column was in MATLAB's datenum²⁰ format, as well as the fact that when data was converted to Excel it appeared in exponential form, thus losing precision. The end result is daily Excel files in the format shown in Figure 19:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	263410880	1	3	2017	558	37,13287	-8,526638333	2	0	221,7	398,9									
2	224769000	1	3	2017	558	42,942655	-9,1804	0	0,1	169,4	429,6									
3	224237920	1	3	2017	558	36,595925	-4,513265	0	0	116,2	38,9									
4	224022920	1	3	2017	558	37,02339667	-6,926858333	0	3,1	182	398,9									
5	224053280	1	3	2017	558	36,87747833	-7,085405	7	3,7	152,4	398,9									

Figure 19 - Excel daily file.

Due to Excel specifications and limits, namely in what concerns the total number of rows and columns in a worksheet (1,048,576 rows by 16,384 columns (Microsoft, n.d.)), it was necessary to convert the .xlsx files into another format, CSV. In order to do so, it was used a Java Script file converter, as it can be observed in Figure 20.

¹⁹ Excel format is .xlsx.

²⁰ MATLAB's datenum function creates a numeric array to represent date and time.

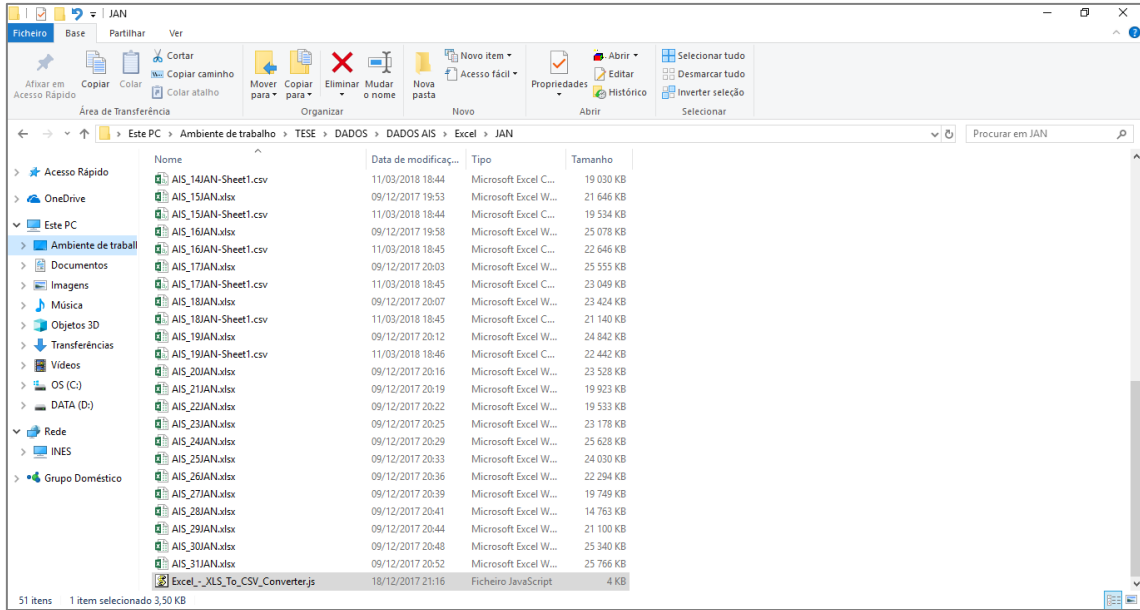


Figure 20 - Conversion from .xlsx files to .csv.

In this new format, it was possible to aggregate the files into monthly .csv files using the command line presented in Figure 21.

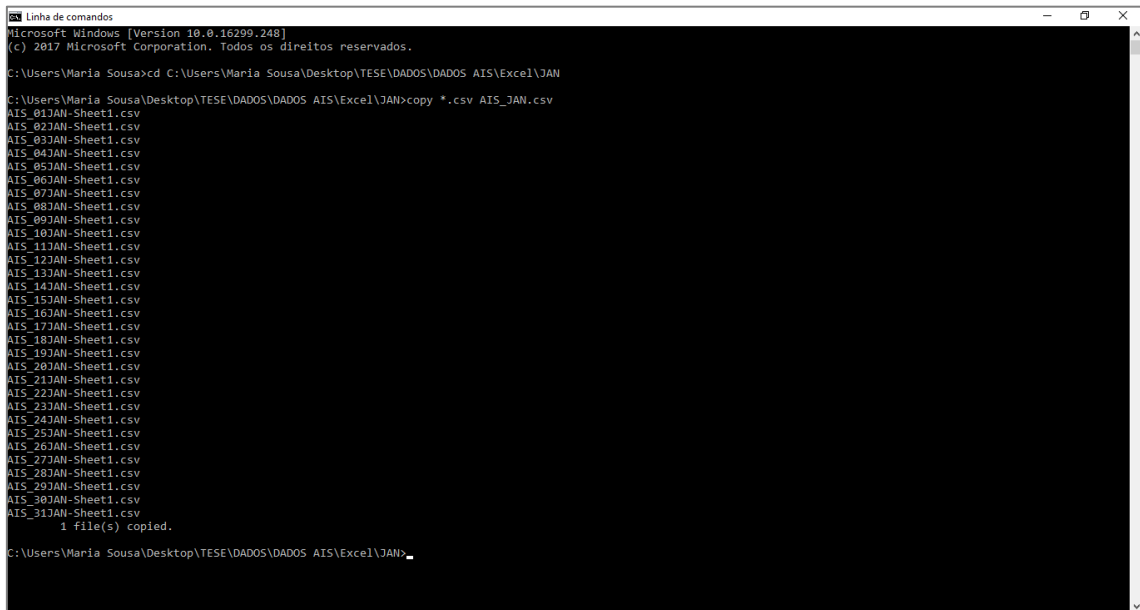


Figure 21 - Aggregation of the daily .csv files into monthly .csv files.

Each AIS monthly file has an average of 11 million rows.

Georeferencing data can be done directly in the database. However, there are other, simpler, ways to do it. For instance, by using ArcMap from ArcGIS, a Geographic Information System (GIS) developed by a company called Esri. ArcMap is “the primary

application used in ArcGIS for Desktop for mapping, editing, analysis, and data management” (ArcGIS, n.d.).

By adding the .csv files as XY data and choosing the World Geodetic System 84 (WGS84), which is the “reference coordinate system used by the Global Positioning System” (GISGeography, 2018), it is possible to save the data in shapefile ²¹ (.shp) format, as represented in Figure 22. According to Esri Shapefile Technical description, a shapefile “stores nontopological geometry and attribute information for the spatial features in a dataset” and the “geometry for a feature is stored as a shape comprising a set of vector coordinates” (Esri, 1997).

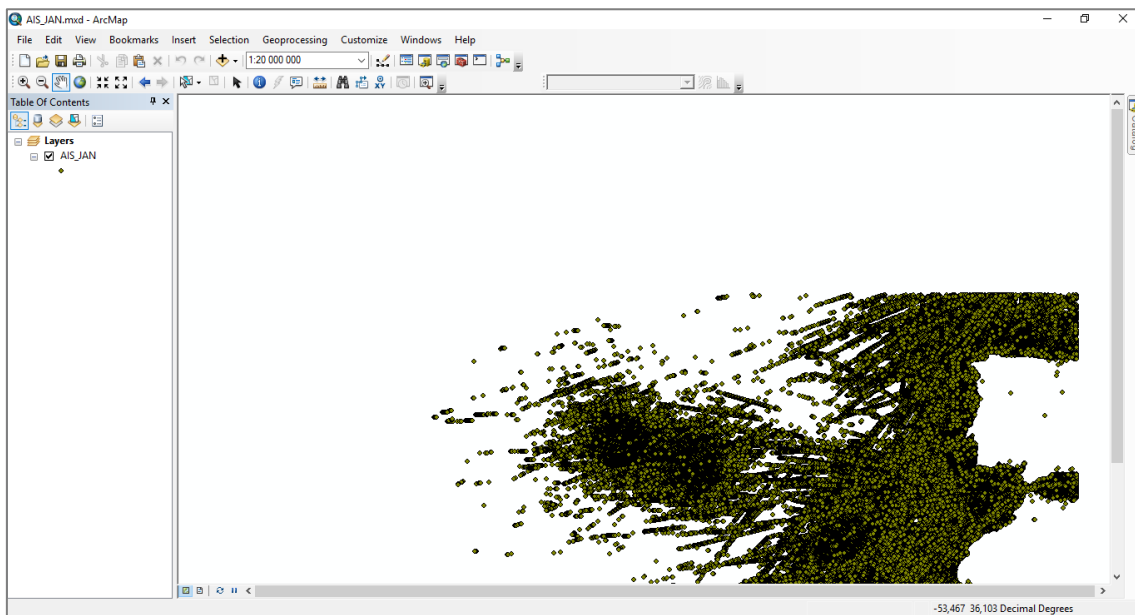


Figure 22 - Shapefile created through ArcGIS.

However, even though ArcGIS stands out by its documentation on how to use its tools, it is, nonetheless, a proprietary software and its licence comes at a cost. For this reason, another software was used throughout this dissertation, QGIS. Through QGIS (which is an open-source cross-platform GIS), it is possible to create and add the shapefiles created as a vector layer, as it is presented in Figure 23.

²¹ A shapefile is an Esri vector data storage format for storing the location, shape, and attributes of geographic features. It is stored as a set of related files and contains one feature class (ArcGIS, n.d.).

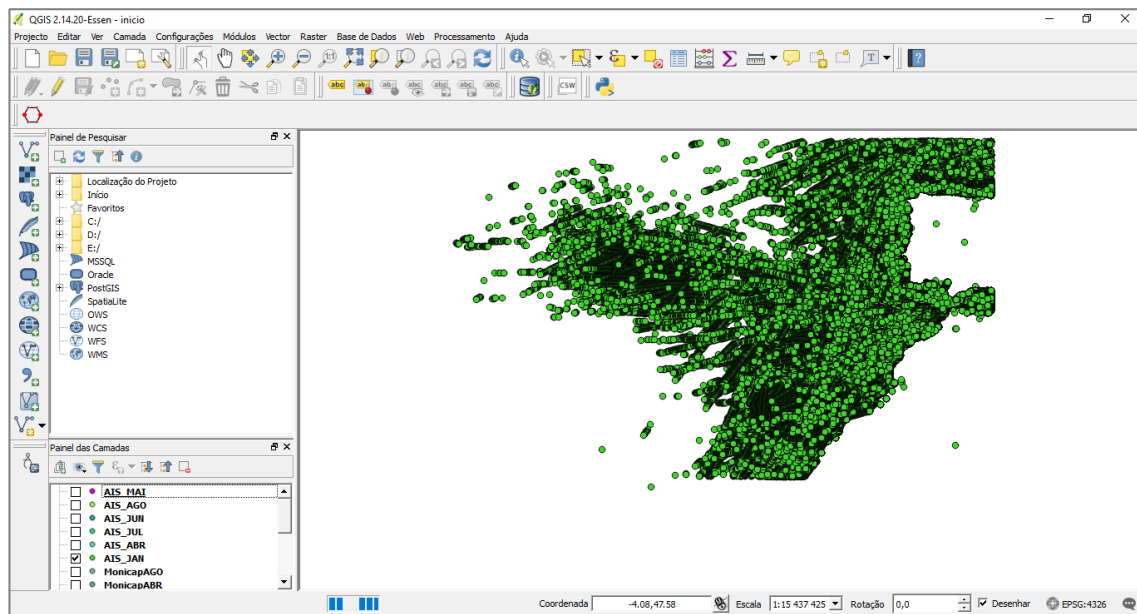


Figure 23 - QGIS vector layer added.

Granted that this step was done successfully, it is then possible to connect PostgreSQL database to QGIS and export that layer (Figure 24).

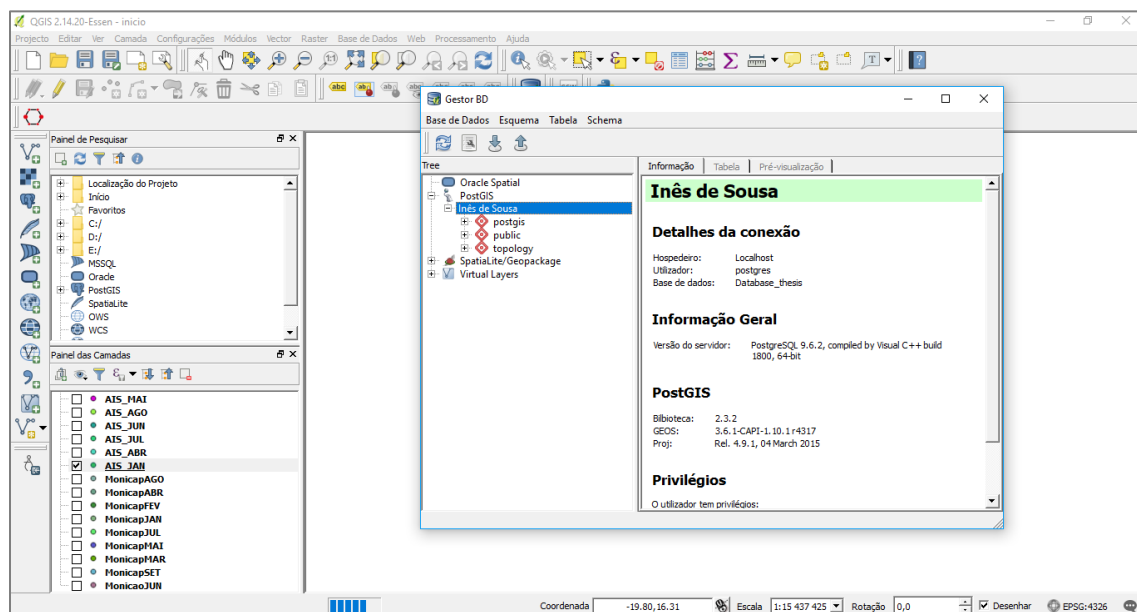


Figure 24 – QGIS to PostgreSQL connection.

Regarding the MONICAP data, the process is equal once the data is in .csv format. One can achieve that by using Power Query, an Excel add-in, to join the different .txt files into one month based .xlsx file. On average, a MONICAP monthly file has about

450 000 rows. Afterwards, the files are converted to .csv and the steps explained above are repeated.

The database management interface pgAdmin was used to manage the database.

Figure 25 shows the database in the pgAdmin interface.

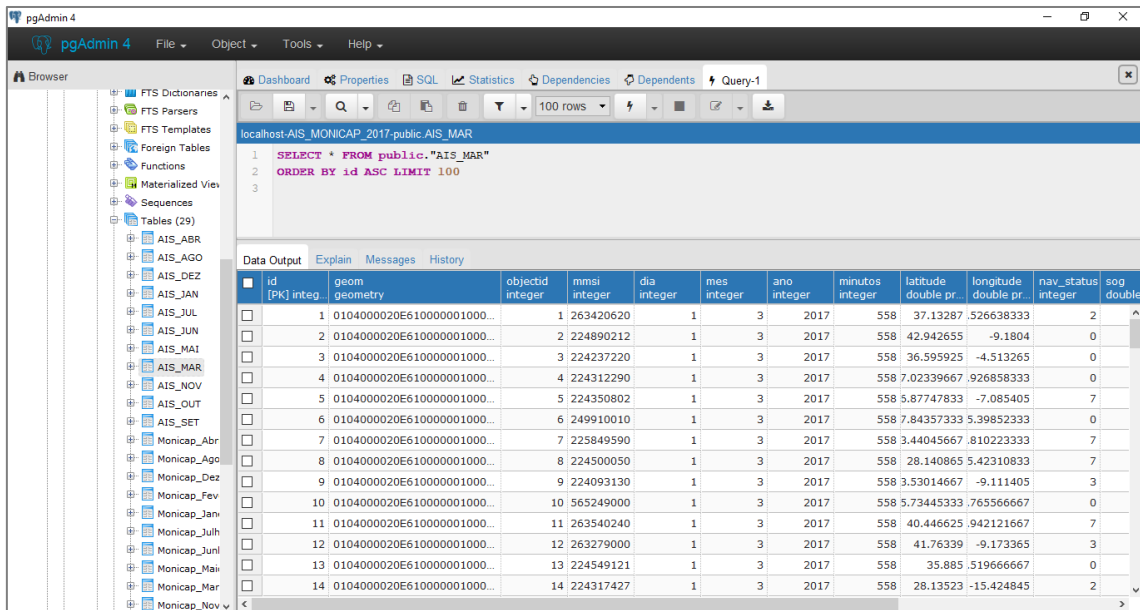


Figure 25 – Database in the pgAdmin interface.

3.4. PostgreSQL Database

There are several reasons that support the choice of PostgreSQL, besides being open source and available for all the most common operating systems as mentioned above. Behind a good documentation, in what concerns manuals, tutorials, books, among others, there is also a very active and collaborative community.

Table 4 represents PostgreSQL processing capability.

Table 4 - PostgreSQL processing capability (Database guide, n.d.).

Limit	Value
Maximum Database Size	Unlimited
Maximum Table Size	32TB
Maximum Row Size	1.6TB
Maximum Field Size	1GB
Maximum Rows per Table	Unlimited

Maximum Columns per Table	250-1600 depending on column types
Maximum Indexes per Table	Unlimited

Another advantage of the usage of PostgreSQL is its advanced database management system, as it allows the access of several users at the same time. In addition, it is also possible to grant different roles for each user. This allows the database to follow the Atomicity, Consistency, Isolation, Durability (ACID) principles that every SQL database guarantee. (Essential SQL, n.d.). Atomicity is the principle where every operation is done as a single unit, which means that, for example, in a transaction moving two data sets, either all the information is saved or none. Consistency, on its turn, means that interrupted changes do not take place. Instead, the process is rolled back in order to keep data integrity. Isolation principle ensures that a certain transaction is not affected by any other transactions taking place. Finally, durability means that once a certain transaction is finished, it will remain so, even in the case of a system failure.

The database is under the name “AIS_MONICAP_2017” and is composed by several tables, each one of them concerning a type of data and a month, as it can be seen in Figure 26.

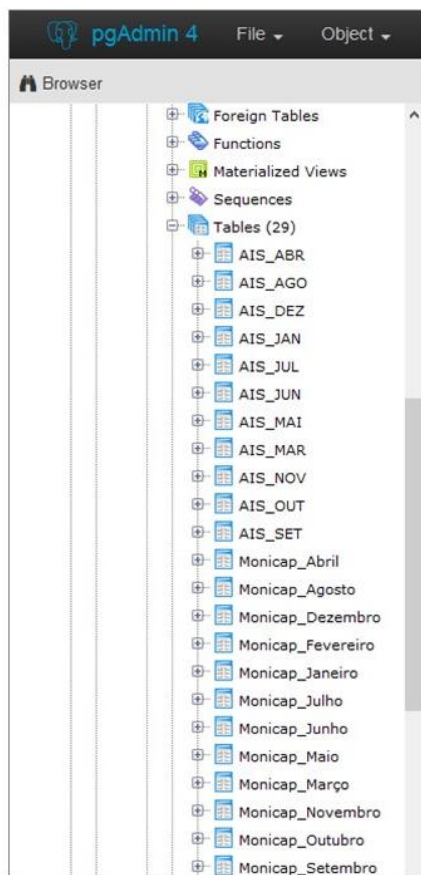


Figure 26 - pgAdmin tables of AIS_MONICAP_2017 database.

Inside each table, the data is organized as shown in Figure 27.

Data Output													Explain	Messages	History
<input type="checkbox"/>	id [PK] int...	geom geometry	objectid integer	mmsi integer	dia integer	mes integer	ano integer	minutos integer	latitude double pr...	longitude double pr...	nav_status integer	sog double pr...	cog double pr...		
<input type="checkbox"/>	1	0104000020E61000000100...	1	263420620	1	3	2017	558	37.13287	.526638333	2	0	221.7		
<input type="checkbox"/>	2	0104000020E61000000100...	2	224890212	1	3	2017	558	42.942655	-9.1804	0	0.1	169.4		
<input type="checkbox"/>	3	0104000020E61000000100...	3	224237220	1	3	2017	558	36.595925	-4.513265	0	0	116.2		
<input type="checkbox"/>	4	0104000020E61000000100...	4	224312290	1	3	2017	558	7.02339667	.926858333	0	3.1	182		
<input type="checkbox"/>	5	0104000020E61000000100...	5	224350802	1	3	2017	558	5.87747833	-7.085405	7	3.7	152.4		
<input type="checkbox"/>	6	0104000020E61000000100...	6	249910010	1	3	2017	558	7.84357333	5.39852333	0	0.1	231.6		
<input type="checkbox"/>	7	0104000020E61000000100...	7	225849590	1	3	2017	558	3.44045667	.810223333	7	2.4	306.5		

Data Output													Explain	Messages	History
<input type="checkbox"/>	id [PK] int...	geom geometry	objectid integer	mmsi integer	dia integer	mes integer	ano integer	minutos integer	latitude double pr...	longitude double pr...	nav_status integer	sog double pr...	cog double pr...		
<input type="checkbox"/>	1	0104000020E61000000100...	1	263420620	1	3	2017	558	37.13287	.526638333	2	0	221.7		
<input type="checkbox"/>	2	0104000020E61000000100...	2	224890212	1	3	2017	558	42.942655	-9.1804	0	0.1	169.4		
<input type="checkbox"/>	3	0104000020E61000000100...	3	224237220	1	3	2017	558	36.595925	-4.513265	0	0	116.2		

Figure 27 - Example of one of the AIS tables created.

One of the first steps was to normalize data in what concerns decimal places. The end result is 5 decimal places for latitude and longitude fields in both AIS and



MONICAP data. The field of SOG will have one decimal place, by reducing one decimal place of MONICAP data to match AIS's field, and COG with 0, to match AIS data to MONICAP's COG field.

The normalizations of SOG and COG are quite easily explained by the irrelevance of 0,01 in terms of speed over ground or 0,1 in terms of course over ground. However, the latitude and longitude decimal places need a little further explaining.

“Degrees of latitude are parallel so, for the most part, the distance between each degree remains constant” (Rosenberg, 2018). The same cannot be applied to longitude degrees, because their distance changes significantly, being farthest apart on top of the line of Equator and converging at the poles.

In the Equator, 1° of latitude corresponds to 60 nautical miles²², and as mentioned above it almost does not vary from the Equator to the poles. As it is, 0,00001° of latitude means approximately 1,11 meters, which is a more than acceptable error in terms of positioning at sea.

Regarding longitude, there is the need to perform a calculation based on the respective latitude. A simple and rough way to do this is by using the formula [1].

1° of Longitude = cosine (latitude in decimal degrees) * length of degree (miles) at equator [1]

A degree of longitude is widest at the equator with a distance of 69.172 miles.

Using [1], to the selected area (latitude 37,5°N to 39,5°N; longitude 010,5°W to 008,5°W), 0,00001° of longitude means approximately between 1,016 meters and 0,989 meters, which is also a satisfactory value.

The main goal of the database is to gather and manage all the data required in order to identify anomalies, not only in the correlation between the different data types, but also anomalous values that the data may present.

Once the database was concluded, a list of possible erroneous records was compiled so as to be looked upon and tagged. Therefore, the following possible erroneous records were listed:

²² One nautical mile is equivalent to 1852 meters.

- Missing records, this means that no information has been received from the sensors;
- Records with positions situated in anomalous places, for example onshore;
- Records with SOGs extremely high, for example over than 100 knots;
- Records with erroneous names, sometimes missing or partially cut.

Once these records were identified, they were marked according to each type of error mentioned above. The decision to tag these records instead of simply removing them can be justified by the fact that what looks useless in most cases might be useful in the detection of anomalies in future studies (Urbano, 2014). In addition to that, sometimes repeated erroneous data can suggest a failure in the source, specifically in the sensors providing the data. Also, a record might have some erroneous values, but the remaining might still be useful and valid to be interpreted.



CHAPTER 4

DATA ANALYSIS TOOLS

- 4.1. Selected area
- 4.2. Data mining for anomaly detection using Orange

4. Data analysis tools

4.1. Selected area

In order to conduct some tests with the data, an area of interest was selected. This area comprises the data located between, in latitude $37,5^{\circ}$ and $39,5^{\circ}\text{N}$ and in longitude, $10,5^{\circ}\text{W}$ and $8,5^{\circ}\text{W}$, being represented in Figure 28.

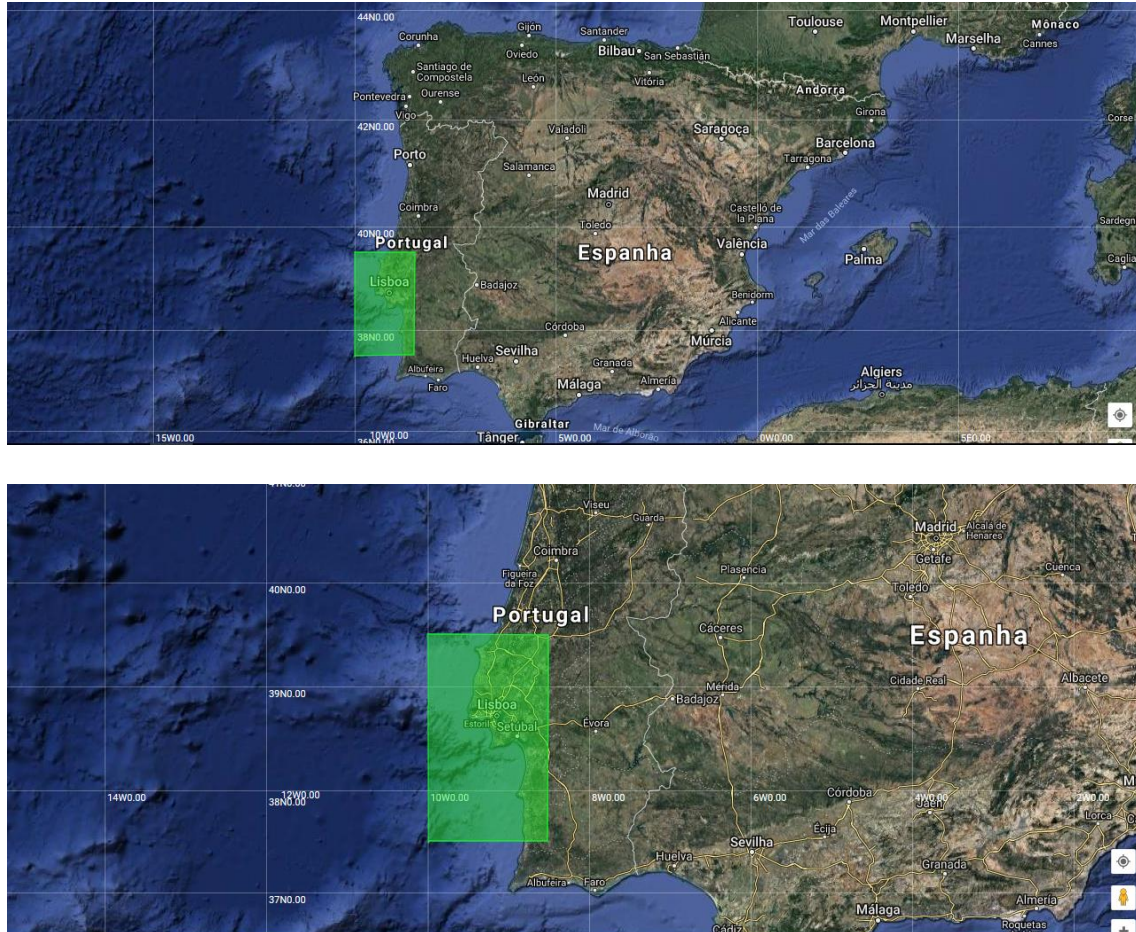


Figure 28 - Selected area (Daftlogic, n.d.).

This area consists of about 8350 Nautical Square Miles (NM^2) and it was chosen through the observation of vessel traffic density near the Portuguese Coast (Figure 29). The main ports comprised in the selected area are Lisbon and Sines. As it was mentioned before, during the introduction, the port of Lisbon stood out in 2017 for its development in terms of cargo movements. The port of Sines, on the other hand, is the port that detains the hegemony of cargo movements in Portugal. The choice of the size of the area was based on having an area large enough to have a comprehensive database without compromising the ability to store and process it.

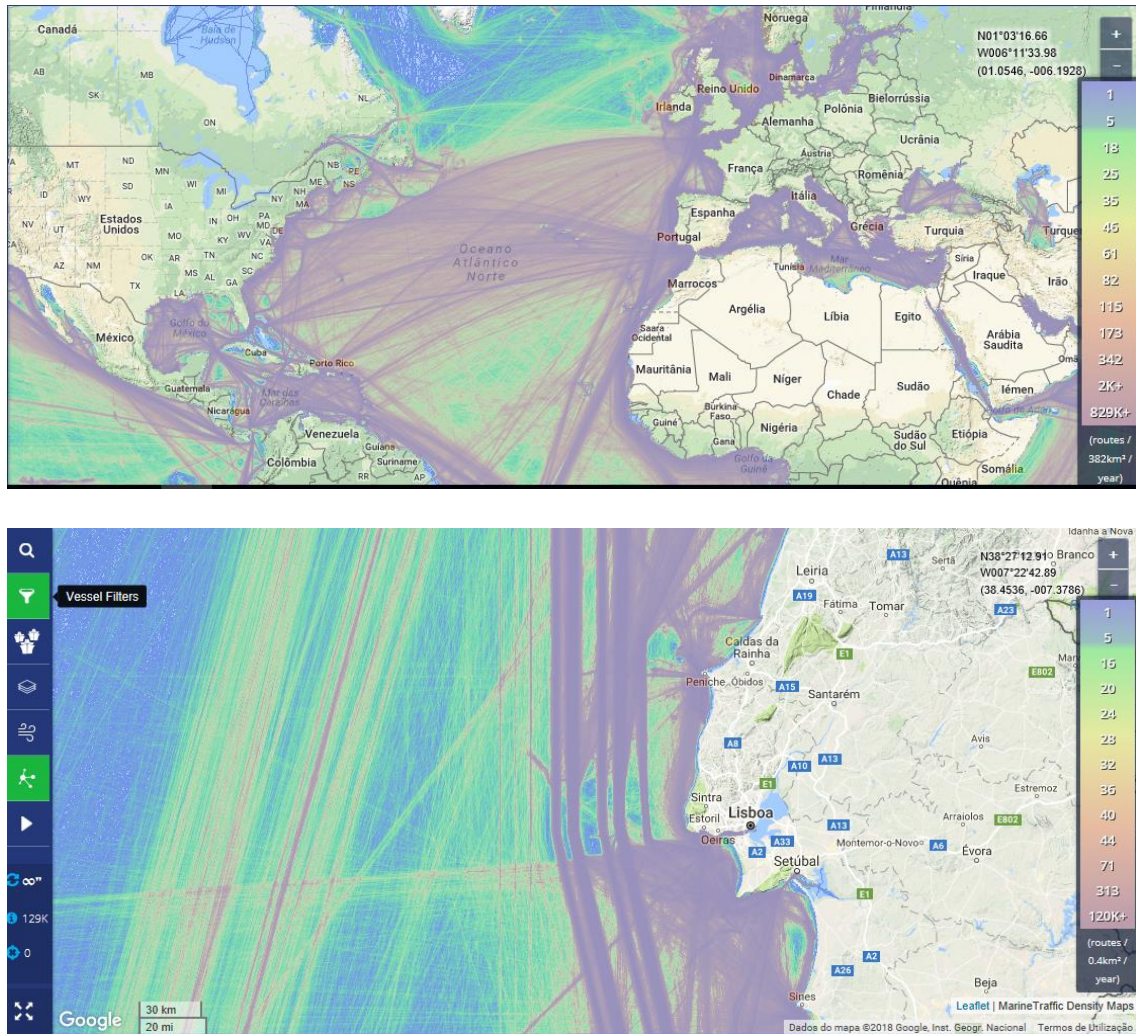


Figure 29 - Traffic density in the chosen area (Marine Traffic, n.d.).

4.2. Data Mining for anomaly detection using Orange

The data mining tool used in the present dissertation was Orange. The criteria was the fact that it is open source, its intuitive workflow and the active developing community, constantly updating and delivering new functionalities.

One of the main advantages of Orange is the ability to connect to a PostgreSQL database, as it is represented in Figure 30. In order to do so, it is necessary to install some extensions, more specifically psycopg2 and quantile.

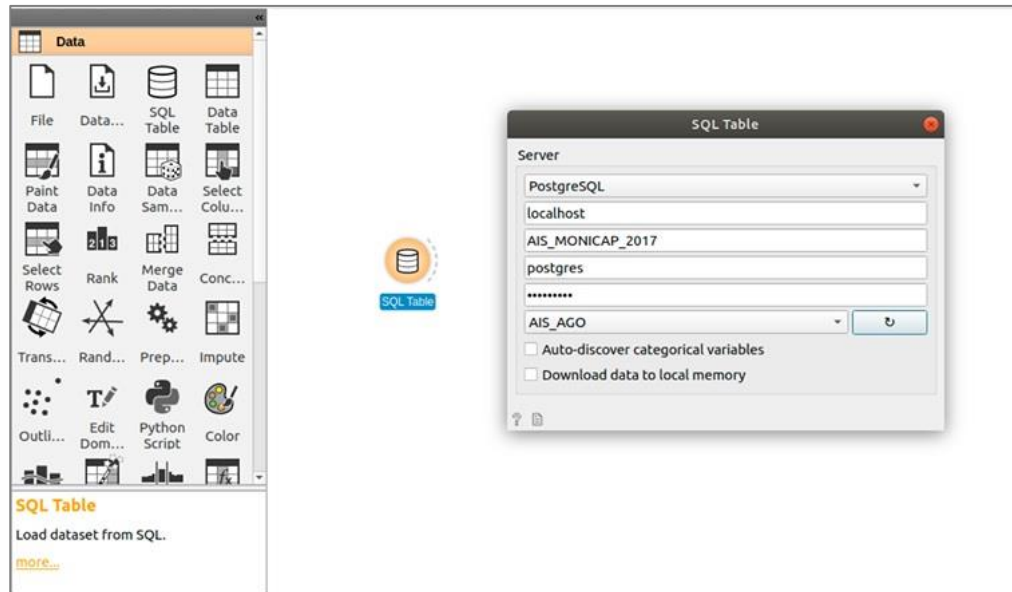


Figure 30 - Orange connection to PostgreSQL database.

With the SQL Table widget it is possible to introduce data from tables of the created database. However, one downside of the usage of this widget is the impossibility to use it with a very large table, which happens in “AIS_MONICAP_2017” database. In order to solve this issue two options were considered. One of them was to use the widget Data Sampler, which would randomly select a subset of data from the input dataset. The second option was to create smaller tables, with less data. Taking into consideration that the main goal was to analyze a certain area and data concerning specific ships, it was not viable to randomly select data. Therefore, smaller tables were created, with a week worth of data, within the selected area.

For the initial trials, it was important to see if there was a simple way to correlate data from the two sources, AIS and MONICAP. The first tool used was the Distance Matrix. A distance matrix shows the distance between types of corresponding data in form of a table. The distances can be calculated through different methods such as Euclidean, Manhattan and Mahalanobis distance, among others, as shown in Figure 31.

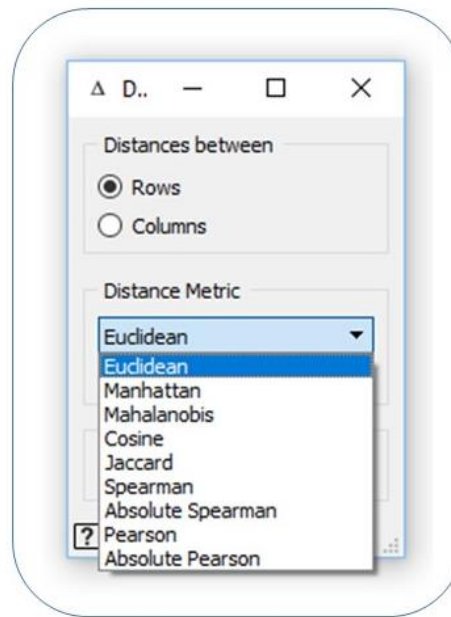


Figure 31 - Distance calculation methods available in Orange.

The second tool used for data correlation was the Impute widget. This widget has the goal of replacing missing values, by computing data according to several techniques or by data set by the user, as it is represented in Figure 32.

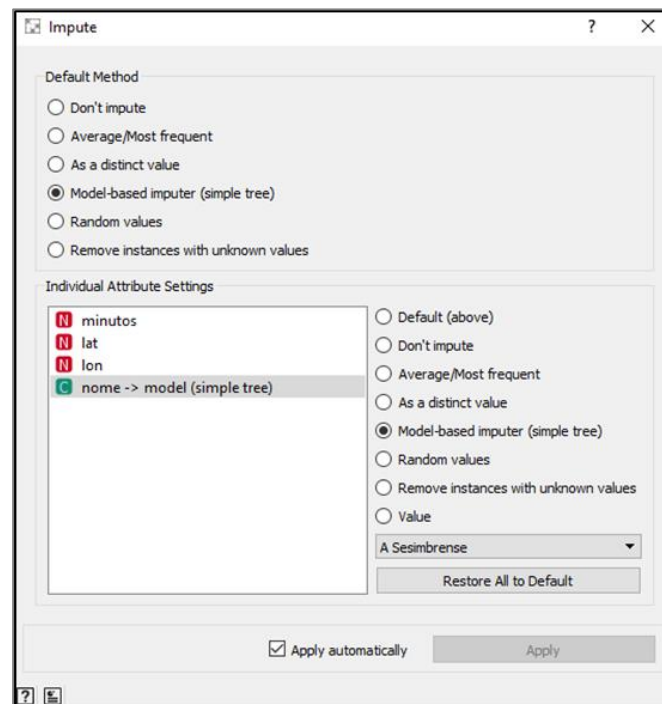


Figure 32 - Orange's impute widget.

It is possible to specify general imputation techniques for every attribute or to specify individual treatment for each attribute. Of all these techniques, the only one that cannot be understood by its term is the Model-based imputer. This technique, by using the values of other attributes, creates a model for predicting the unknown values (Orange, n.d.). The default model is a 1-NN learner that replaces missing values with values from the most similar record. This model derives from a k-Nearest Neighbour (k-NN) algorithm which is used for classification as well as regression-type prediction problems (Sharda, 2014). The k-NN method is one of the simplest of all machine-learning algorithms. For classification prediction, for example, a record is classified according to the class of its k nearest neighbours. According to the values of k , there can be different classifications. In Figure 33, the star in the centre is the target of this classification, being the squares and circles the possible classes. If $k=1$, the star would be considered a square. However, if $k=3$, the classification would be a circle and if $k=5$ it would be a square again. Figure 33 shows the importance of the k value.

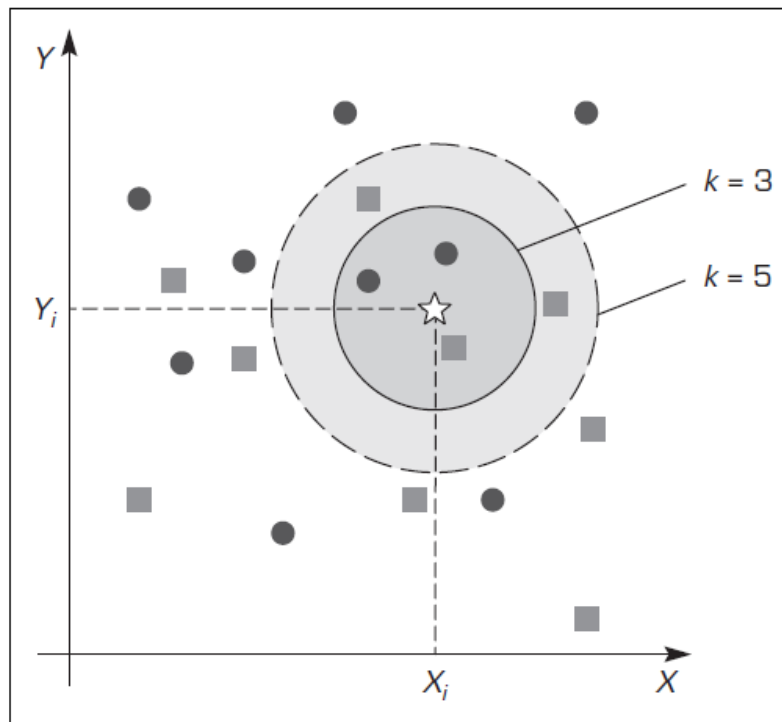


Figure 33 - Different classification based on k value (Sharda, 2014).

In Orange's impute case in particular, 1-NN, the record is classified with the class of its nearest neighbour.

Outlier detection was another tool used during the experimental phase. Orange offers this widget with two different outlier detection methods: One class Support Vector

Machine (SVM) with non-linear kernel and Covariance estimator. While one class SVM with non-linear kernels works on non-Gaussian distributions, Covariance estimator is only used with Gaussian distributions. A Gaussian or normal distribution is a continuous probability distribution, represented by the mean (μ), the standard deviation (σ) and the variance (σ^2). This distribution can be observed in Figure 34. As the data used in this dissertation does not follow a Gaussian distribution, one class SVM with non-linear kernel was used instead (Orange, n.d.).

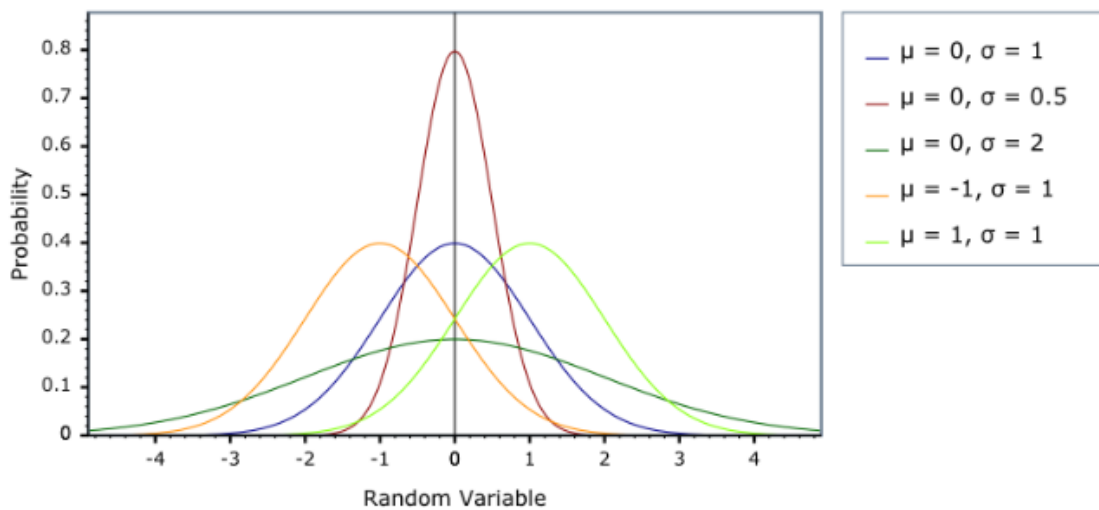


Figure 34 - Example of Gaussian distribution (Boost, n.d.).

Learning algorithms of data mining methods can be classified as supervised or unsupervised. Unsupervised learning is used when a model is trained without labelled data. On the other hand, supervised machine learning happens when labelled training data is available. One class SVM with non-linear kernel is an unsupervised learning algorithm that classifies data as similar or different from the training set (Sharda, 2014).

Using the Outliers widget in Orange (Figure 35), it is possible to define two parameters for the One class SVM with non-linear kernel method: Nu and Kernel coefficient. Nu parameter is an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors. Kernel coefficient is a gamma parameter, which determines the influence of a single data record. As this algorithm does not have an optimal set of parameters by default, it is necessary to experiment with these parameters in order to optimise the results obtained (Orange, n.d.).

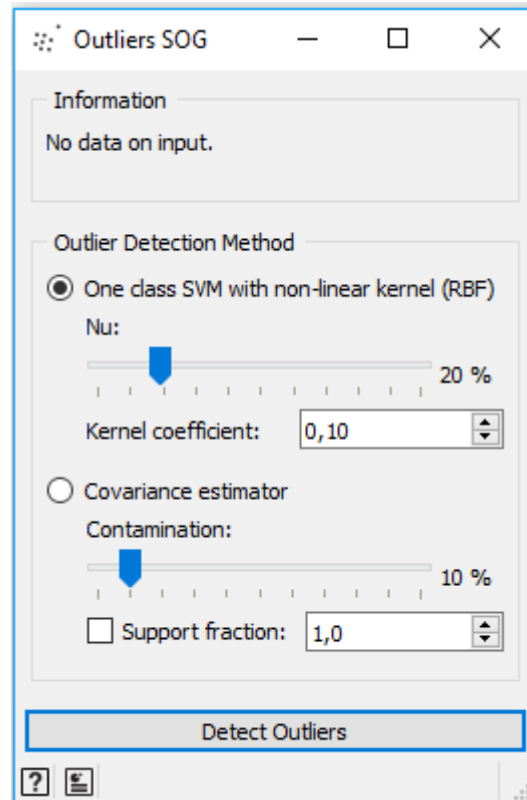


Figure 35 - Orange's outlier widget.



CHAPTER 5

TESTS AND RESULTS

- 5.1 Data correlation tests and results
- 5.2 Anomaly detection tests and results

5. Tests and Results

5.1. Data correlation tests and results

5.1.1. Test 1

In the first test, AIS and MONICAP data that corresponded to a single ship's route, *Aurora*, was used. This route concerns data from one day, 1st of March 2017. This first test uses a .csv file, *rotas_correspondentes*, with one vessel, having the complete AIS and MONICAP records from that day.

As it was said before, Orange is quite simple to use. The first step is to insert the file with the data required through the file widget (Figure 36).

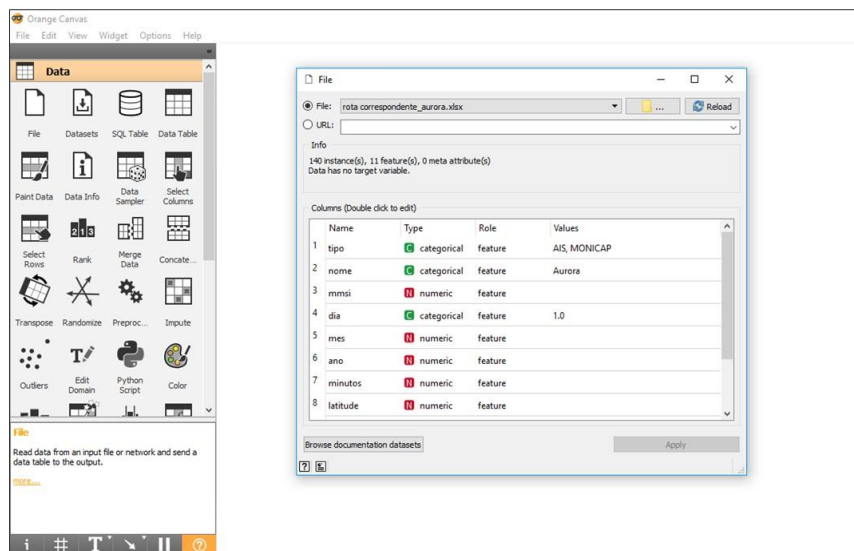


Figure 36 - Data file insertion.

Afterwards, it is necessary to select the features considered relevant to enter the distance matrix through the select columns widget. In this test, the features selected were minutes, latitude and longitude. These features can be selected in the select column widget, represented in Figure 37. With it, it is possible to select the features that will be used for the intended algorithm, so as the features that will be selected as targets and as meta attributes, which are features that, for the most cases, are not considered in the analysis (Orange, n.d.).

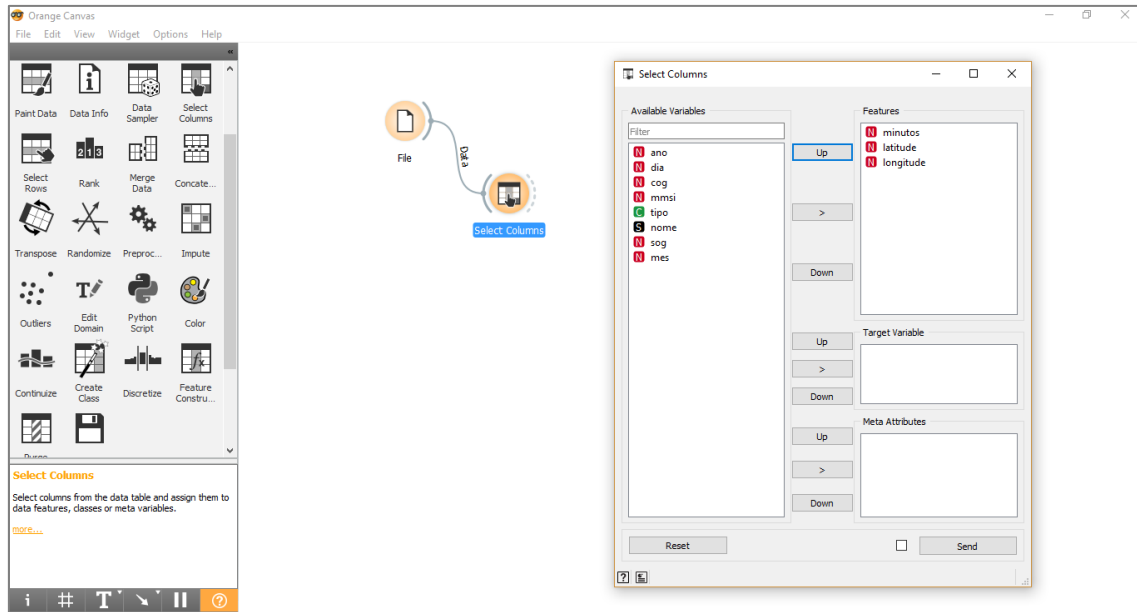


Figure 37 - Selection of features.

Then, it is necessary to apply the distances widget and choose the type of distance wanted in the distance matrix. In this situation, Euclidean distance was the chosen method because, considering that the data is from a relatively small area, it could be compared to calculating distances in a two-dimensional plan (Figure 38).

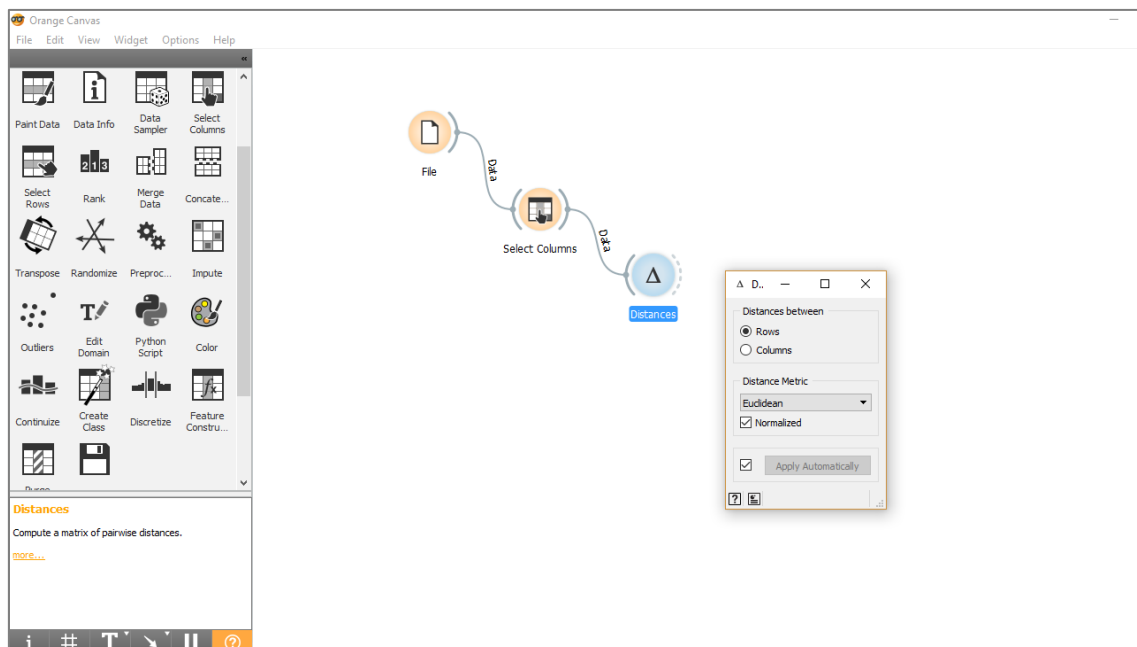


Figure 38 - Distance selected, Euclidean.

Only later it is possible to apply the distance matrix widget to the data, as it is represented in Figure 39.

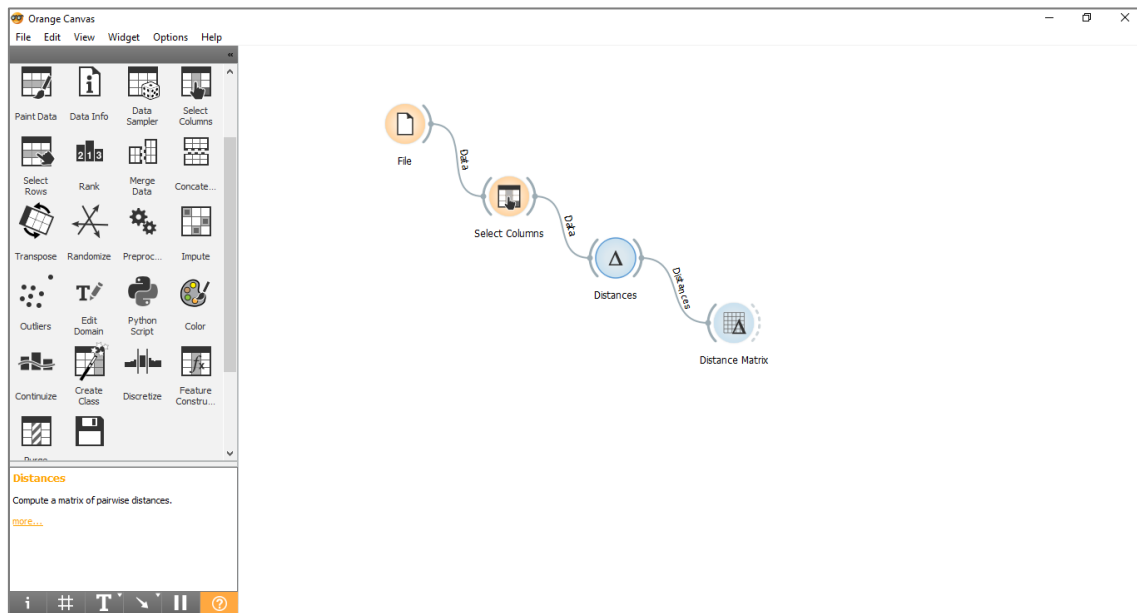


Figure 39 - Distance matrix application.

Distance matrix results appear in a continuous color gradient ranging from a darker green, high values, to white, values equal or close to 0, as it can be observed in Figure 40.

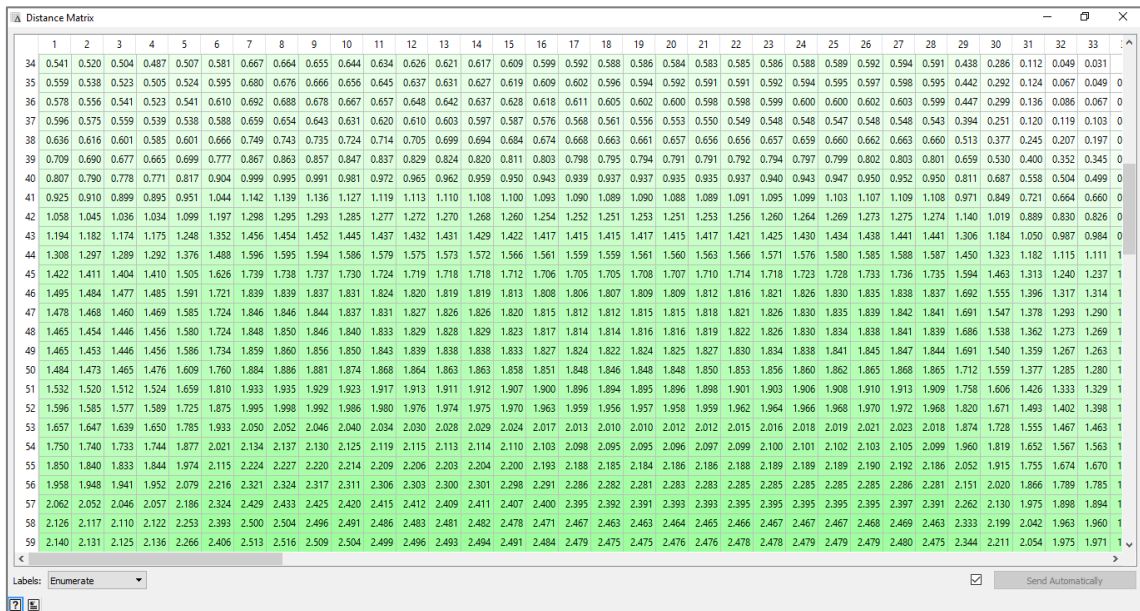
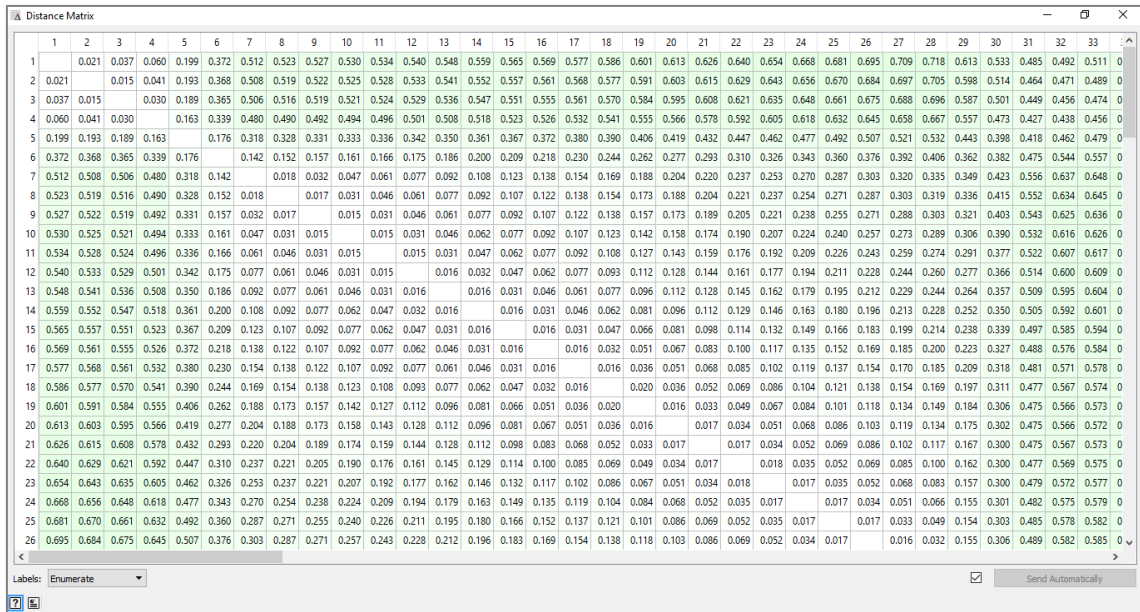


Figure 40 – Example of colour gradient based on distances.

AIS data goes from row 1 to 99, MONICAP data goes from row 100 to 141, in file *rotas_correspondentes*, all ordered in function of the time.

From Figure 41, it is possible to see that there are big distances between the first points in AIS and the final ones, which makes sense because the ship is moving, from a point A to a Point B that are not the same, and consequently is altering its position. The time feature also contributes to the disparity of the values.

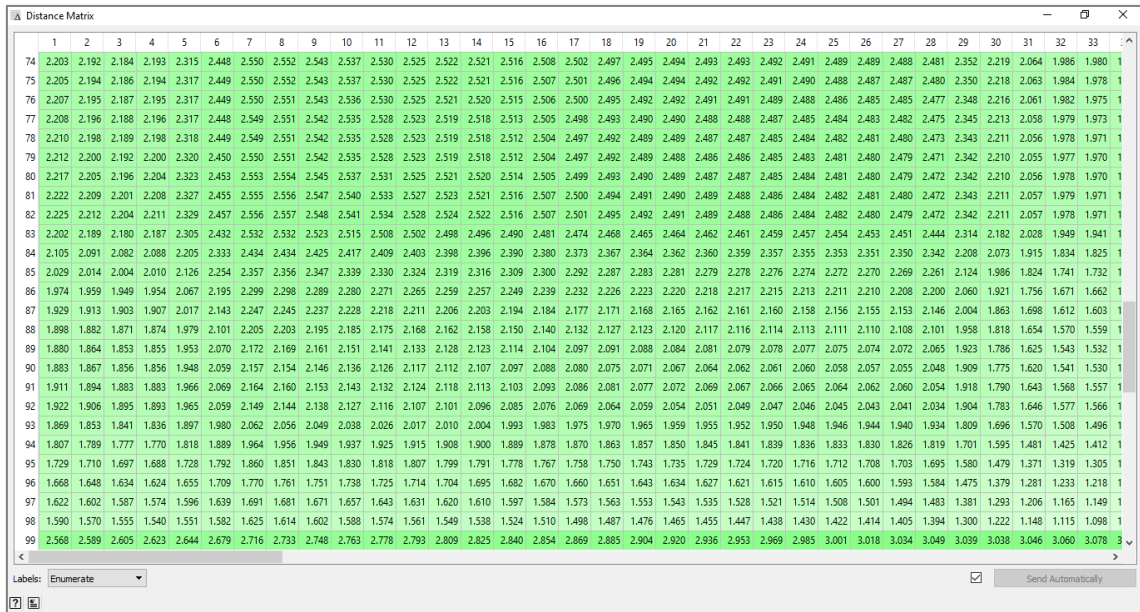
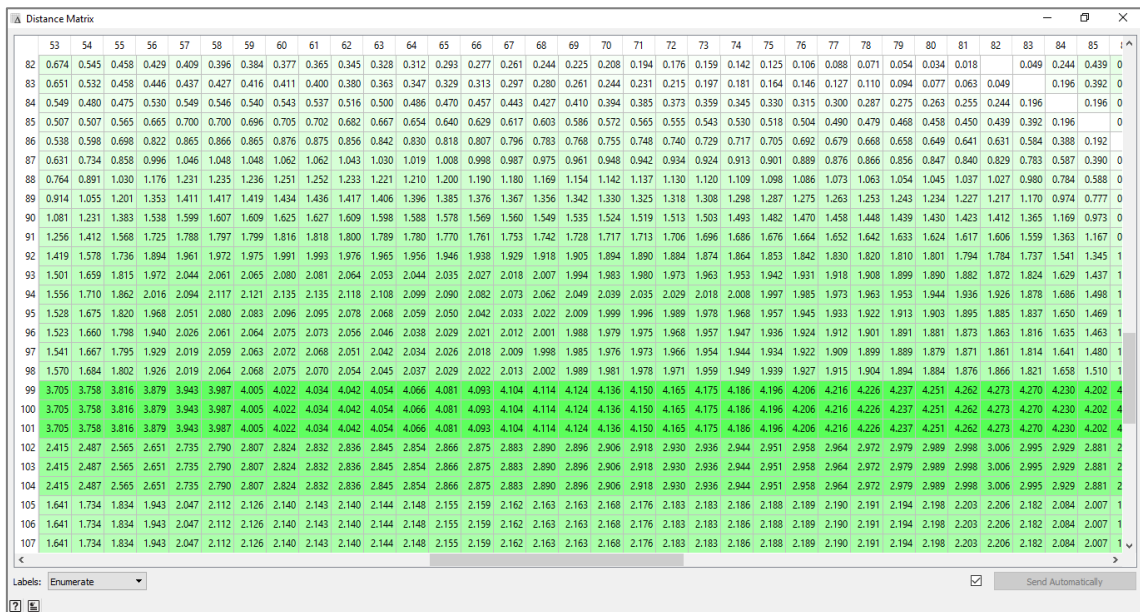
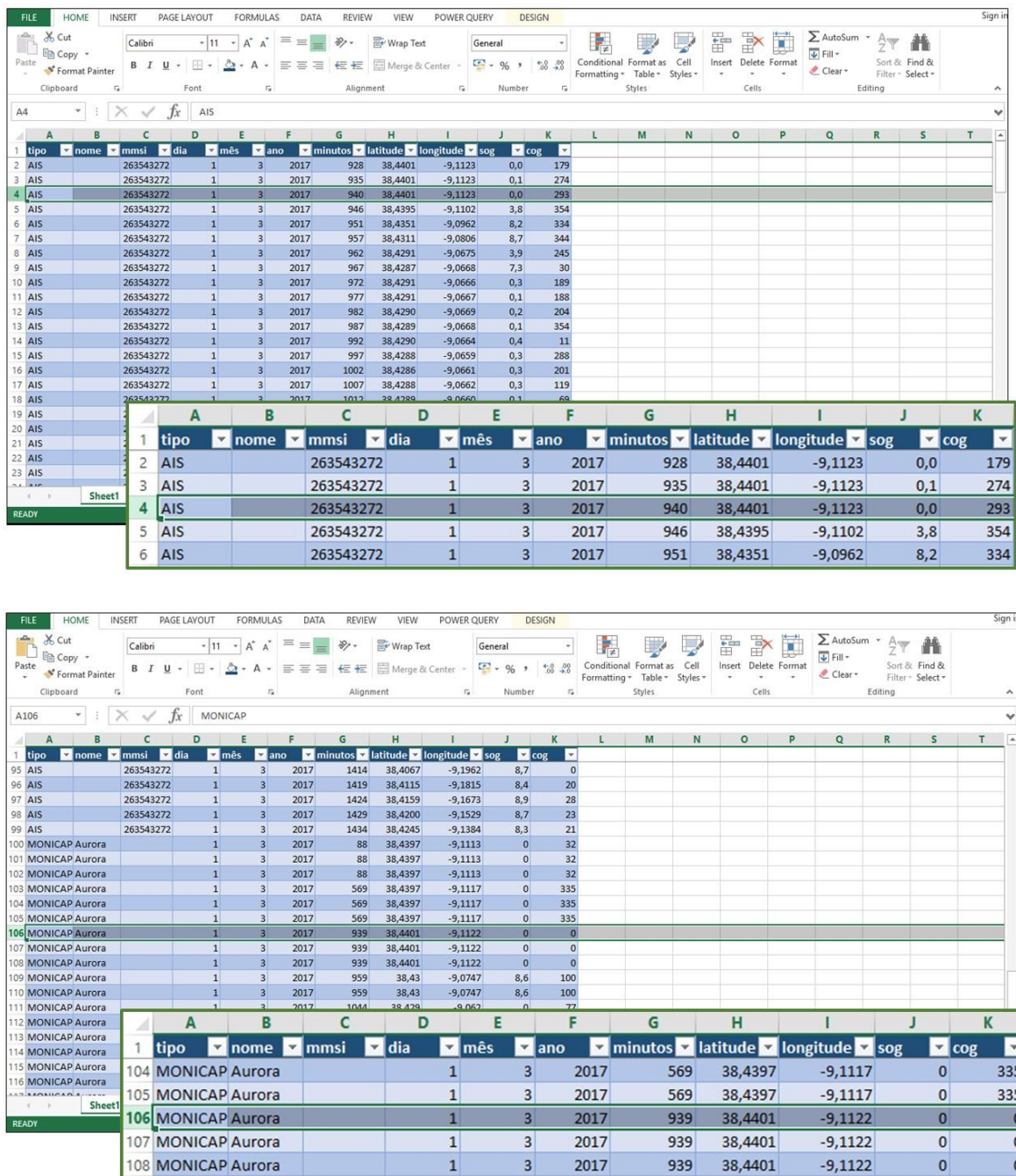


Figure 41 - Distance matrix results between AIS data.

The biggest differences observed in this matrix (Figure 42) is when comparing the final AIS data with the first reports of MONICAP, which also makes sense considering that it is not a circular route.



Because the file used is one that we can be assured that belongs to the same vessel, it is possible to select data (AIS and MONICAP) from a certain instance in time and see if the correspondent distance shows a small value (indexes are different because of the heading of the Excel file). It was selected data from almost the same time period, with only a minute of difference (Figure 43).



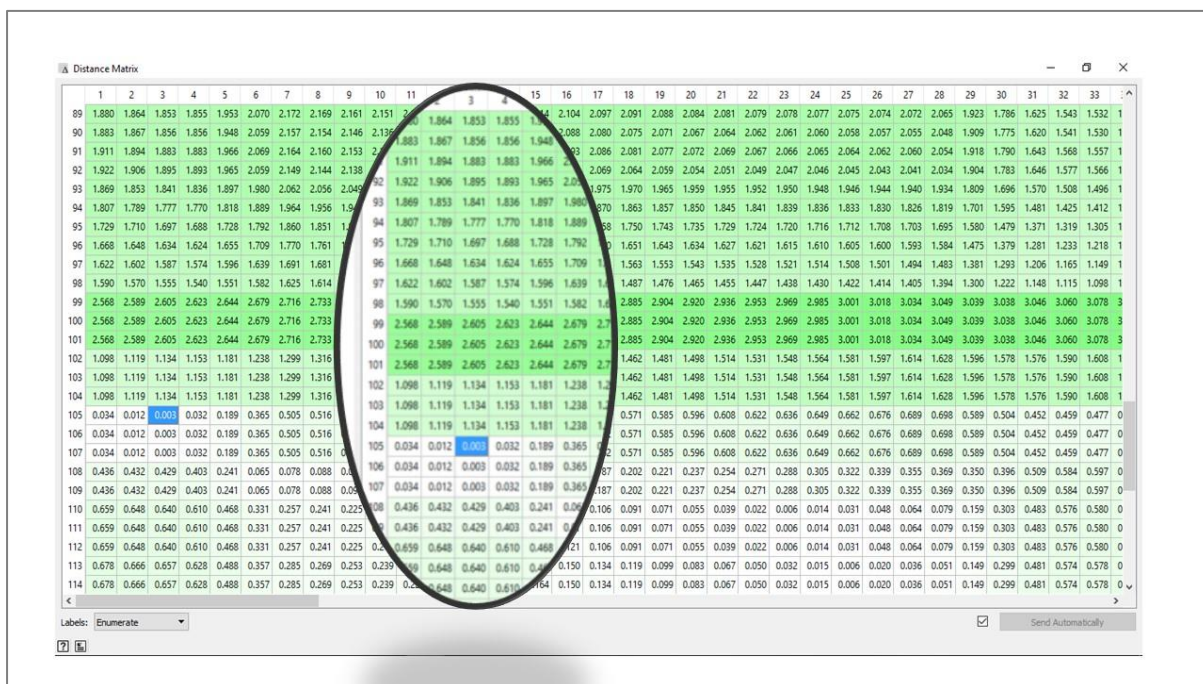
tipo	nome	mmsi	dia	mês	ano	minutos	latitude	longitude	sog	cog
1										
2	AIS	263543272	1	3	2017	928	38,4401	-9,1123	0,0	179
3	AIS	263543272	1	3	2017	935	38,4401	-9,1123	0,1	274
4	AIS	263543272	1	3	2017	940	38,4401	-9,1123	0,0	293
5	AIS	263543272	1	3	2017	946	38,4395	-9,1102	3,8	354
6	AIS	263543272	1	3	2017	951	38,4351	-9,0962	8,2	334
7	AIS	263543272	1	3	2017	957	38,4311	-9,0806	8,7	344
8	AIS	263543272	1	3	2017	962	38,4291	-9,0675	3,9	245
9	AIS	263543272	1	3	2017	967	38,4287	-9,0668	7,3	30
10	AIS	263543272	1	3	2017	972	38,4291	-9,0666	0,3	189
11	AIS	263543272	1	3	2017	977	38,4291	-9,0667	0,1	188
12	AIS	263543272	1	3	2017	982	38,4290	-9,0669	0,2	204
13	AIS	263543272	1	3	2017	987	38,4289	-9,0668	0,1	354
14	AIS	263543272	1	3	2017	992	38,4290	-9,0664	0,4	11
15	AIS	263543272	1	3	2017	997	38,4288	-9,0659	0,3	288
16	AIS	263543272	1	3	2017	1002	38,4286	-9,0661	0,3	201
17	AIS	263543272	1	3	2017	1007	38,4288	-9,0662	0,3	119
18	AIS	263543272	1	3	2017	1012	38,4288	-9,0660	0,1	69
19	AIS									
20	AIS									
21	AIS									
22	AIS									
23	AIS									

tipo	nome	mmsi	dia	mês	ano	minutos	latitude	longitude	sog	cog
1										
2	AIS	263543272	1	3	2017	928	38,4401	-9,1123	0,0	179
3	AIS	263543272	1	3	2017	935	38,4401	-9,1123	0,1	274
4	AIS	263543272	1	3	2017	940	38,4401	-9,1123	0,0	293
5	AIS	263543272	1	3	2017	946	38,4395	-9,1102	3,8	354
6	AIS	263543272	1	3	2017	951	38,4351	-9,0962	8,2	334

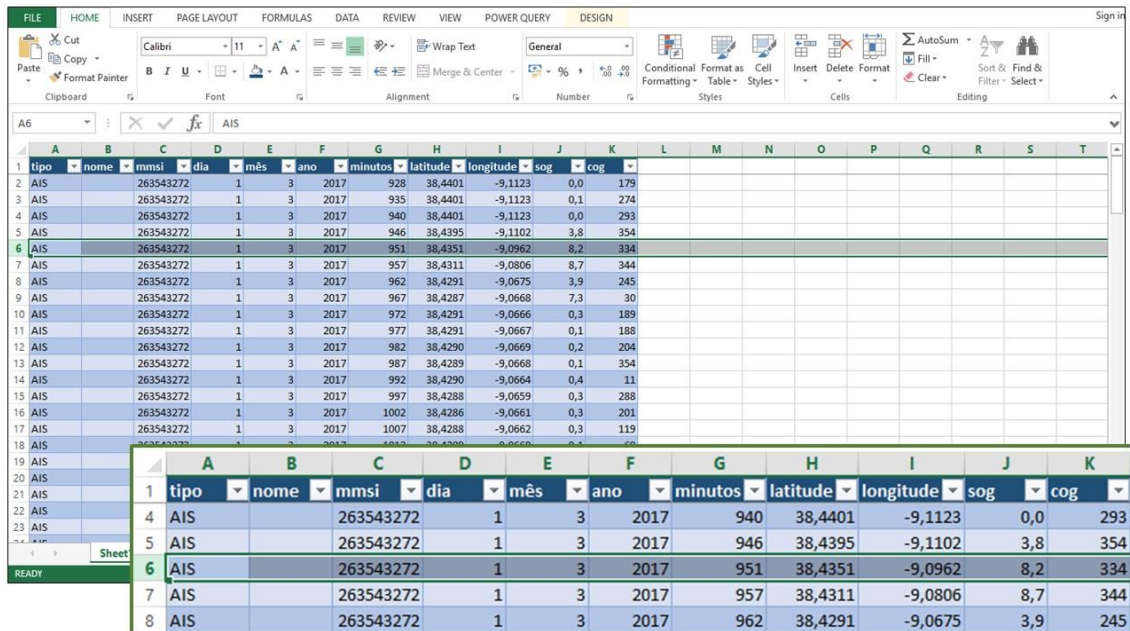
tipo	nome	mmsi	dia	mês	ano	minutos	latitude	longitude	sog	cog
1										
95	AIS	263543272	1	3	2017	1414	38,4067	-9,1962	8,7	0
96	AIS	263543272	1	3	2017	1419	38,4115	-9,1815	8,4	20
97	AIS	263543272	1	3	2017	1424	38,4159	-9,1673	8,9	28
98	AIS	263543272	1	3	2017	1429	38,4200	-9,1529	8,7	29
99	AIS	263543272	1	3	2017	1434	38,4245	-9,1384	8,3	21
100	MONICAP Aurora		1	3	2017	88	38,4397	-9,1113	0	32
101	MONICAP Aurora		1	3	2017	88	38,4397	-9,1113	0	32
102	MONICAP Aurora		1	3	2017	88	38,4397	-9,1113	0	32
103	MONICAP Aurora		1	3	2017	569	38,4397	-9,1117	0	335
104	MONICAP Aurora		1	3	2017	569	38,4397	-9,1117	0	335
105	MONICAP Aurora		1	3	2017	569	38,4397	-9,1117	0	335
106	MONICAP Aurora		1	3	2017	939	38,4401	-9,1122	0	0
107	MONICAP Aurora		1	3	2017	939	38,4401	-9,1122	0	0
108	MONICAP Aurora		1	3	2017	939	38,4401	-9,1122	0	0
109	MONICAP Aurora		1	3	2017	959	38,43	-9,0747	8,6	100
110	MONICAP Aurora		1	3	2017	959	38,43	-9,0747	8,6	100
111	MONICAP Aurora		1	3	2017	1044	38,428	-9,062	0	77
112	MONICAP Aurora									
113	MONICAP Aurora									
114	MONICAP Aurora									
115	MONICAP Aurora									
116	MONICAP Aurora									

Figure 43 - Values used to analyse results of distance matrix.

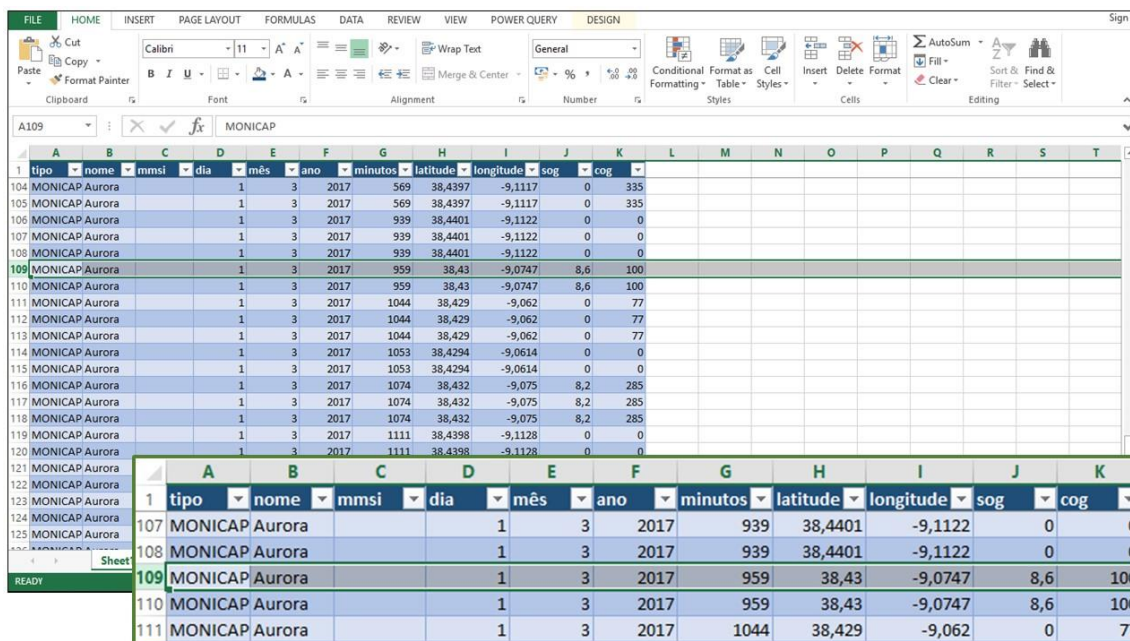
The value obtained was 0.003, close to 0, as it is represented in Figure 44.



Afterwards, the goal was to assess the impact that the time feature had on the results of the distance matrix. Therefore, it was selected two points with an 8 minute time difference (Figure 45).



1	tipo	nome	mmsi	dia	mês	ano	minutos	latitude	longitude	sog	cog
2	AIS		263543272	1	3	2017	928	38,4401	-9,1123	0,0	179
3	AIS		263543272	1	3	2017	935	38,4401	-9,1123	0,1	274
4	AIS		263543272	1	3	2017	940	38,4401	-9,1123	0,0	293
5	AIS		263543272	1	3	2017	946	38,4395	-9,1102	3,8	354
6	AIS		263543272	1	3	2017	951	38,4351	-9,0962	8,2	334
7	AIS		263543272	1	3	2017	957	38,4311	-9,0806	8,7	344
8	AIS		263543272	1	3	2017	962	38,4291	-9,0675	3,9	245



1	tipo	nome	mmsi	dia	mês	ano	minutos	latitude	longitude	sog	cog
104	MONICAP	Aurora		1	3	2017	569	38,4397	-9,1117	0	335
105	MONICAP	Aurora		1	3	2017	569	38,4397	-9,1117	0	335
106	MONICAP	Aurora		1	3	2017	939	38,4401	-9,1122	0	0
107	MONICAP	Aurora		1	3	2017	939	38,4401	-9,1122	0	0
108	MONICAP	Aurora		1	3	2017	939	38,4401	-9,1122	0	0
109	MONICAP	Aurora		1	3	2017	959	38,43	-9,0747	8,6	100
110	MONICAP	Aurora		1	3	2017	959	38,43	-9,0747	8,6	100
111	MONICAP	Aurora		1	3	2017	1044	38,429	-9,062	0	77

Figure 45 – Values used to analyze distance matrix with 8 minutes difference.

The result was a 0.403, quite different from the previous result (Figure 46).

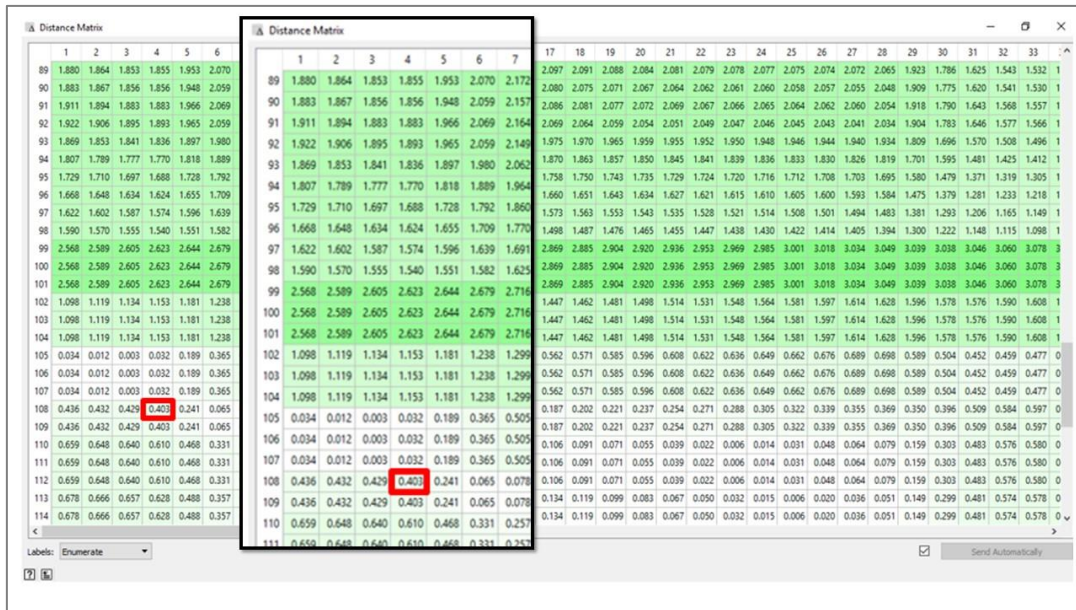


Figure 46 - Results obtained from correlating AIS and MONICAP data.

5.1.2. Test 2

In test 2, the data used is the same from the previous test. The difference is that time is not chosen as a feature, and therefore, it will not be considered in the computations (Figure 47). The results are then compared with the ones from test 1.

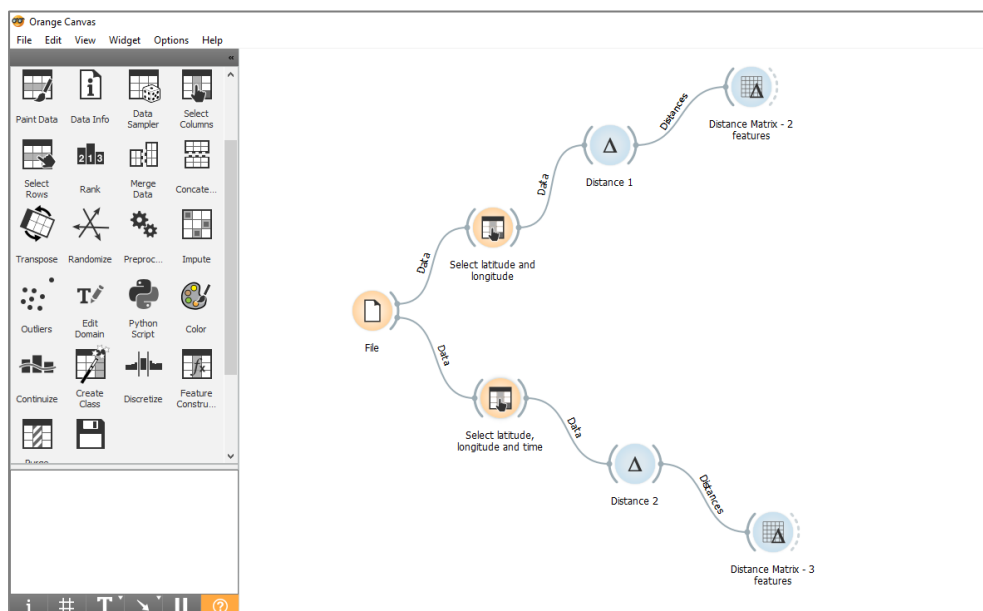


Figure 47 - Orange scheme without time data.

At first sight there are already differences (Figure 48).

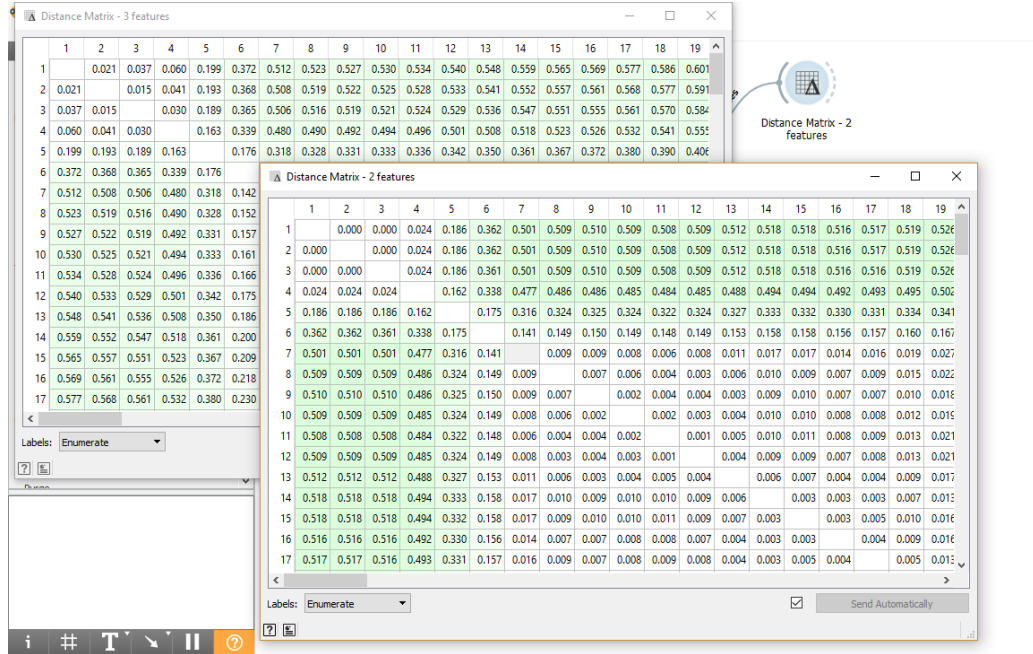


Figure 48 - Distance matrix results without time data.

With the 3 features, the overall distance is bigger than with only latitude and longitude but there is not a big discrepancy. The first MONICAP points start at 100 and in the matrix on the left the distances are very big because time didn't correspond in the two different data (Figure 49).

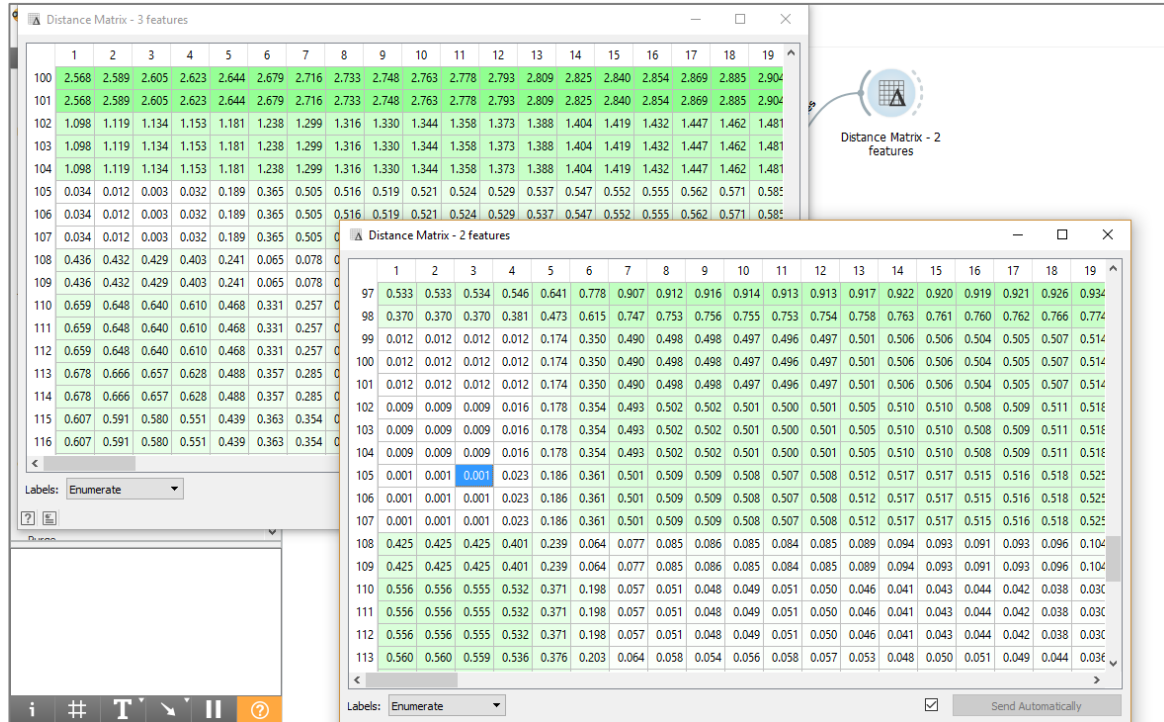


Figure 49 - Distance matrix results for test 2.

5.1.3. Test 3

a) Latitude, Longitude and Time features

The main goal was to identify a correspondence between single ship's route with AIS and MONICAP data with 3 other ship's route with only MONICAP data, all concerning one day of data (1st of March 2017) with latitude, longitude and time as features. Also, during test 3, the applicability of the Impute widget in this dataset was verified by confirming if it is possible to use it to replace missing values, and therefore to associate MMSIs from AIS data with names from MONICAP data.

The file used throughout this test has four different MONICAP vessels and one AIS, existing one correspondence between one MMSI and one name.

In this test, despite the distances regarding the corresponding MONICAP not being constant, it is only with the correspondent MONICAP data that the values get so close to 0. It is also possible to see that, in spite of a 5 minute difference between the points presented below, the distance is of 0.000 which makes sense by looking at the data (Figure 50) and seeing that the ship at that point has a SOG equal to 0.

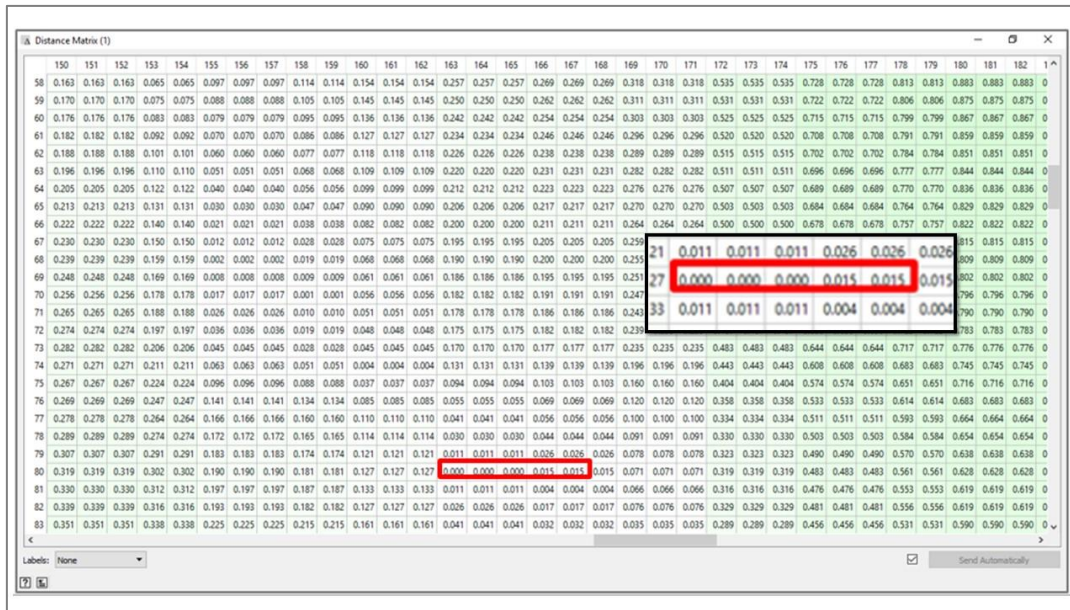
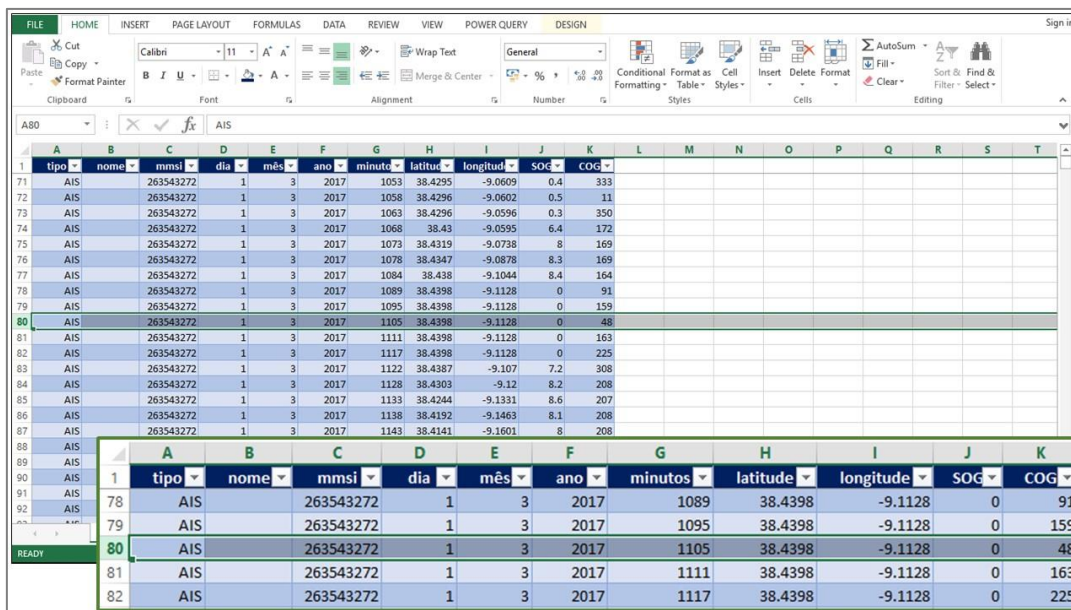
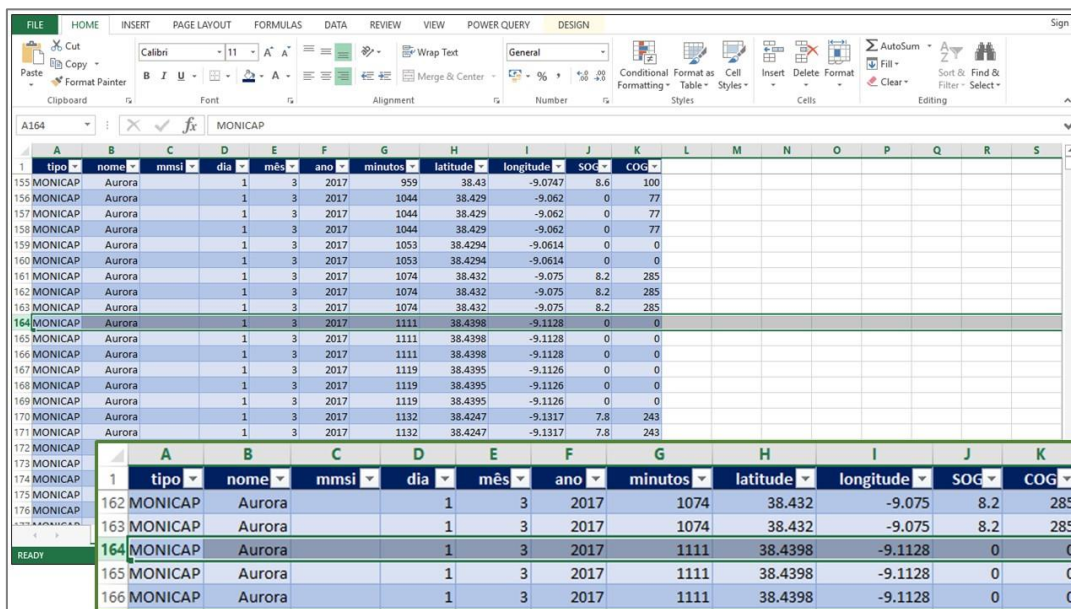


Figure 50 - Distance matrix results in test 3.

Figure 51 represents a sample of the data used for this test.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	tipo	nome	mmsi	dia	mês	ano	minutos	latitude	longitude	SOG	COG									
71	AIS	263543272	1	3	2017	1053	38.4295	-9.0609	0.4	333										
72	AIS	263543272	1	3	2017	1058	38.4296	-9.0602	0.5	11										
73	AIS	263543272	1	3	2017	1063	38.4296	-9.0596	0.3	350										
74	AIS	263543272	1	3	2017	1068	38.43	-9.0595	6.4	172										
75	AIS	263543272	1	3	2017	1073	38.4319	-9.0738	8	169										
76	AIS	263543272	1	3	2017	1078	38.4347	-9.0878	8.3	169										
77	AIS	263543272	1	3	2017	1084	38.438	-9.1044	8.4	164										
78	AIS	263543272	1	3	2017	1089	38.4398	-9.1128	0	91										
79	AIS	263543272	1	3	2017	1095	38.4398	-9.1128	0	159										
80	AIS	263543272	1	3	2017	1105	38.4398	-9.1128	0	48										
81	AIS	263543272	1	3	2017	1111	38.4398	-9.1128	0	163										
82	AIS	263543272	1	3	2017	1117	38.4398	-9.1128	0	225										
83	AIS	263543272	1	3	2017	1122	38.4387	-9.107	7.2	308										
84	AIS	263543272	1	3	2017	1128	38.4303	-9.12	8.2	208										
85	AIS	263543272	1	3	2017	1133	38.4244	-9.1331	8.6	207										
86	AIS	263543272	1	3	2017	1138	38.4192	-9.1463	8.1	208										
87	AIS	263543272	1	3	2017	1143	38.4141	-9.1601	8	208										



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	tipo	nome	mmsi	dia	mês	ano	minutos	latitude	longitude	SOG	COG								
155	MONICAP	Aurora	1	3	2017	959	38.43	-9.0747	8.6	100									
156	MONICAP	Aurora	1	3	2017	1044	38.429	-9.062	0	77									
157	MONICAP	Aurora	1	3	2017	1044	38.429	-9.062	0	77									
158	MONICAP	Aurora	1	3	2017	1044	38.429	-9.062	0	77									
159	MONICAP	Aurora	1	3	2017	1053	38.4294	-9.0614	0	0									
160	MONICAP	Aurora	1	3	2017	1053	38.4294	-9.0614	0	0									
161	MONICAP	Aurora	1	3	2017	1074	38.432	-9.075	8.2	285									
162	MONICAP	Aurora	1	3	2017	1074	38.432	-9.075	8.2	285									
163	MONICAP	Aurora	1	3	2017	1074	38.432	-9.075	8.2	285									
164	MONICAP	Aurora	1	3	2017	1111	38.4398	-9.1128	0	0									
165	MONICAP	Aurora	1	3	2017	1111	38.4398	-9.1128	0	0									
166	MONICAP	Aurora	1	3	2017	1111	38.4398	-9.1128	0	0									
167	MONICAP	Aurora	1	3	2017	1119	38.4395	-9.1126	0	0									
168	MONICAP	Aurora	1	3	2017	1119	38.4395	-9.1126	0	0									
169	MONICAP	Aurora	1	3	2017	1119	38.4395	-9.1126	0	0									
170	MONICAP	Aurora	1	3	2017	1132	38.4247	-9.1317	7.8	243									
171	MONICAP	Aurora	1	3	2017	1132	38.4247	-9.1317	7.8	243									

Figure 51 - AIS and MONICAP data used on test 3.

For the Impute widget, latitude, longitude and time were selected as features, while name was selected as target, as it is the object of the imputation. Finally, the results obtained from the Impute widget proved to be successful, as the AIS vessel obtained the corresponding name after this process. This can be verified in Figure 52.

Info		nome	mmsi	tipo	latitude	longitude	↕	
267 instances		85	Aurora	263543272	AIS	38.4099	-9.2199	154
3 features (no missing values)		86	Aurora	263543272	AIS	38.4166	-9.2312	125
Discrete class with 4 values (no missing values)		87	Aurora	263543272	AIS	38.4282	-9.2324	92
2 meta attributes (31.6% missing values)		88	Aurora	263543272	AIS	38.4394	-9.2314	88
		89	Aurora	263543272	AIS	38.4493	-9.2305	269
		90	Aurora	263543272	AIS	38.4371	-9.2301	273
Variables		91	Aurora	263543272	AIS	38.4250	-9.2299	274
<input checked="" type="checkbox"/> Show variable labels (if present)		92	Aurora	263543272	AIS	38.4139	-9.2253	305
<input checked="" type="checkbox"/> Visualize numeric values		93	Aurora	263543272	AIS	38.4088	-9.2116	345
<input type="checkbox"/> Color by instance classes		94	Aurora	263543272	AIS	38.4067	-9.1962	0
		95	Aurora	263543272	AIS	38.4115	-9.1815	20
Selection		96	Aurora	263543272	AIS			
<input checked="" type="checkbox"/> Select full rows		97	Aurora	263543272	AIS			
		98	Aurora	263543272	AIS			
		99	Hydra	?	MONICAP			
		100	Hydra	?	MONICAP			
		101	Hydra	?	MONICAP			
		102	Hydra	?	MONICAP			
		103	Hydra	?	MONICAP			
		104	Hydra	?	MONICAP			
		105	Hydra	?	MONICAP			
		106	Hydra	?	MONICAP			
		<						
Restore Original Order								
<input checked="" type="checkbox"/> Send Automatically								

Figure 52 - Results after Impute widget.

b) Latitude, longitude and COG features

In this test, the file used was the same as on the previous one, but the features selected were latitude, longitude and COG.

It turns out to be inconclusive, as data that should be corresponding (AIS: 46 to 143; MONICAP: 144 to 186) appear with high values (Figure 53). In spite of this, the impute widget still presented correct results as it classified the AIS record with the right name.

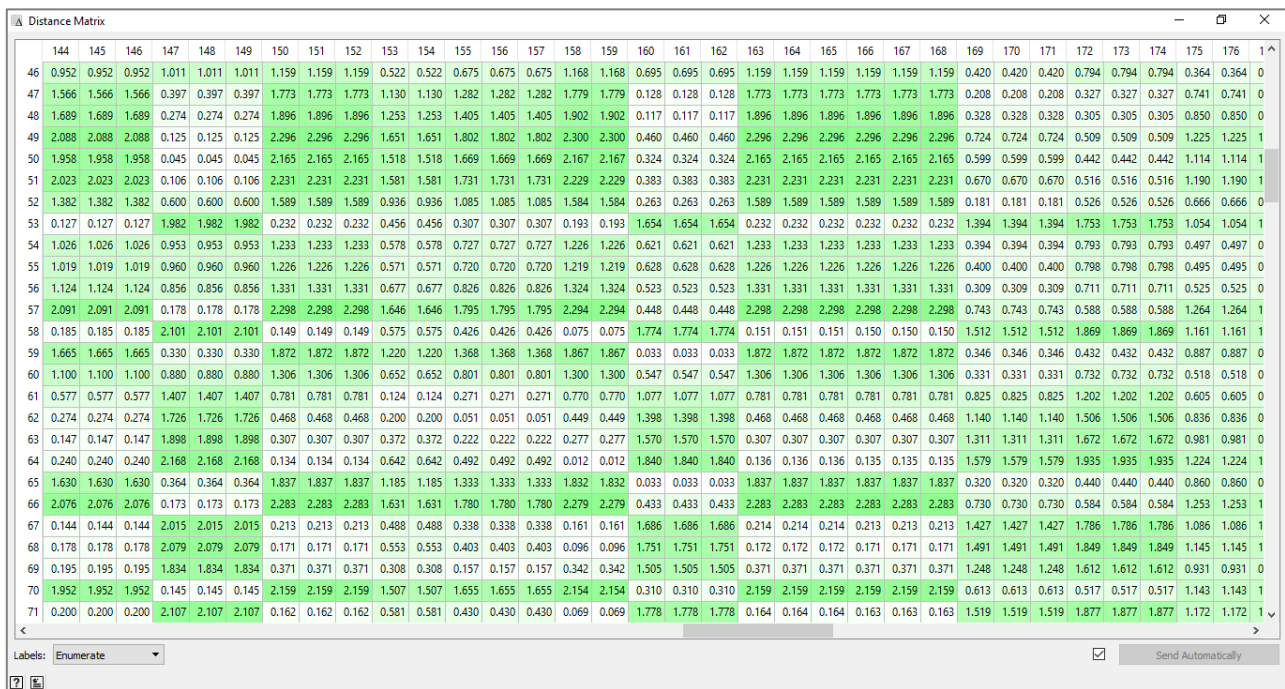


Figure 53 - Distance matrix results in test 3 b).

5.2. Anomaly Detection tests and results

The second orange scenario consists in applying simple filters in order to detect anomalies, for example in terms of speed, and comparing those results with the ones obtained from the orange outlier widget. This scenario uses data from the created database, using the SQL widget in Orange. More specifically, it is connected to a table specially created for Orange tests, *ais_monicap_mar*. This table contains AIS and MONICAP data from the first week of March, as it reduced Orange's processing time. In addition to that, a filter was applied in order to select a specific time for the analysis, in this case, from 0600 to 1200.

This Orange workflow is represented in Figure 54, and it is essentially divided in three parts: detecting vessels with a SOG equal to zero, detecting vessels with anomalous SOGs and detecting vessels with anomalous SOGs or COGs in Cape Roca Traffic Separation Scheme (TSS)²³. It also allows the analysis of a specific vessel that may have been detected as anomalous.

²³ A Traffic Separation Scheme has the main goal of regulating vessels traffic and ensuring navigational safety in congested waters.

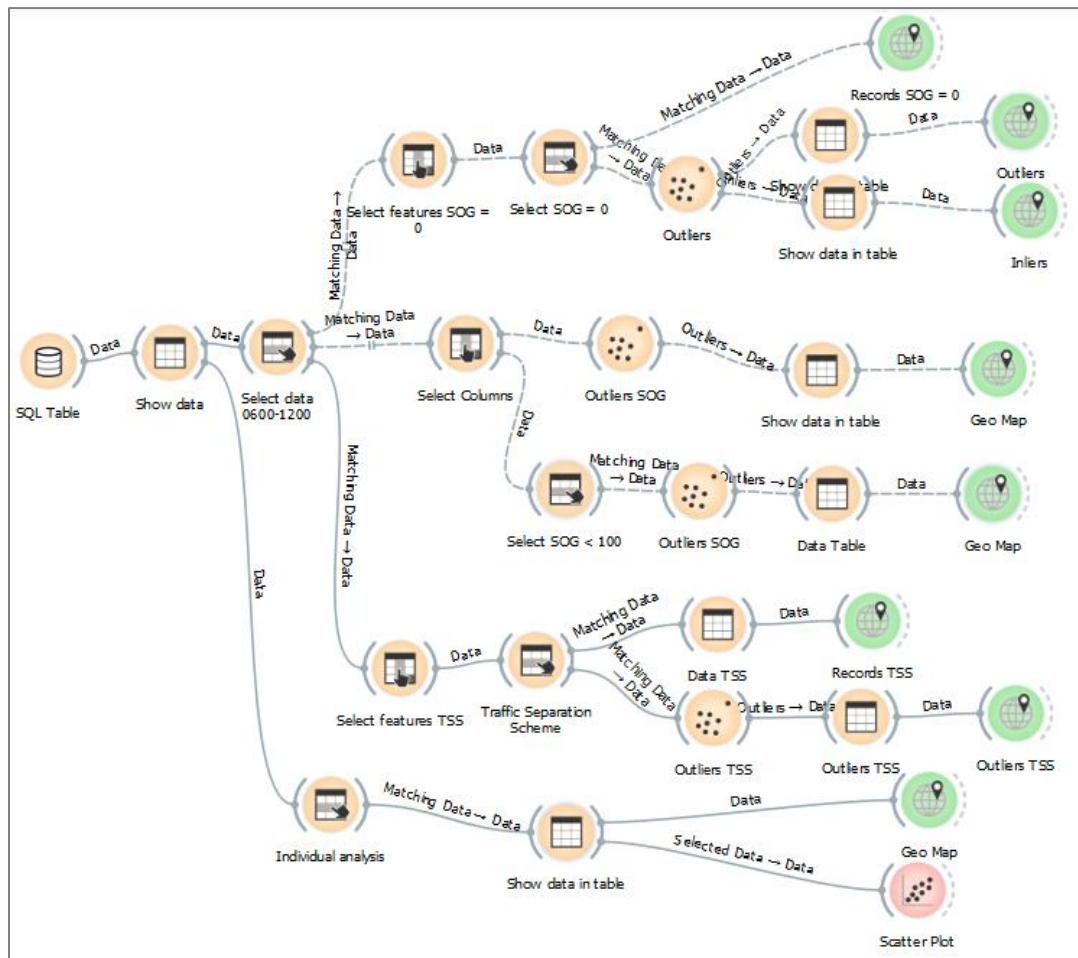


Figure 54 - Anomaly detection test Orange Workflow.

The great advantage of this workflow is that it allows a common user, for example an operator, to change the parameters of what he wants to see without being required any special learning process. The results obtained in each part of the workflow can be represented in a data table, but also in a Geo Map widget.

However, by using the SQL widget, it was not possible to select MMSI and name data as categorical type, as the type of data was automatically chosen from the PostgreSQL table. As a consequence, the impute widget did not present valid results, as it is possible to observe in Figure 55. This happens because MMSI can take any value by being numeric, which would not happen if this feature was set as categorical. Therefore, instead of doing a classification prediction type, it is doing a regression prediction, resulting in MMSI with continuous values and non-existing ones. As a consequence, this widget will not be used in the test.

40562	263500208.000	Pacific	MONICAP	38.520	-8.898	1282.000
40563	263415243.034	Usurper	MONICAP	39.356	-9.368	1121.000
40564	263415243.034	The Intrepid	MONICAP	39.355	-9.368	1120.000
40565	247256400.000	Bermuda	MONICAP	37.574	-8.991	1042.000
40566	263441655.742	Bermuda	MONICAP	37.568	-8.949	1068.000
40567	636014656.000	Bermuda	MONICAP	37.536	-8.924	1311.000
40568	204284992.000	The Duchess	MONICAP	39.355	-9.374	1229.000
40569	258146321.067	Raven	MONICAP	39.298	-9.518	1013.000
40570	258146321.067	Raven	MONICAP	39.294	-9.527	1017.000
40571	263364566.204	Raven	MONICAP	39.273	-9.603	1137.000
40572	263364566.204	Raven	MONICAP	39.207	-9.618	1257.000
40573	263406384.000	Neptune	MONICAP	39.232	-9.468	1012.000
40574	263406384.000	Neptune	MONICAP	39.272	-9.465	1072.000
40575	263379008.000	Neptune	MONICAP	39.336	-9.387	1117.000
40576	263415243.034	Neptune	MONICAP	39.355	-9.372	1132.000
40577	263416448.000	Hydra	MONICAP	38.438	-9.113	1052.000

Figure 55 - Impute widget erroneous results.

For the first segment of the workflow, an analysis is done based on the vessels' SOG being equal to zero, as it is represented in Figure 56. It is also possible to analyze just one type of data (AIS or MONICAP), selecting it in the select data widget.

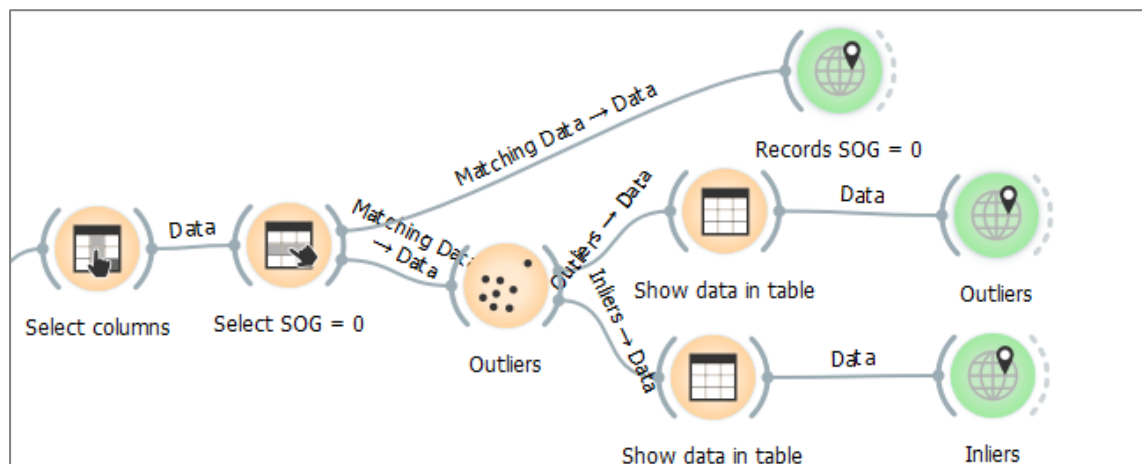


Figure 56 - Branch of the workflow - Analysing records SOG = 0.

As one could expect, most of the records with SOG equal to zero represent moored vessels. However, using the Geo Map widget, it is possible to detect vessels with this speed at sea, which can be observed in Figure 57. This Figure represents every record that has SOG equal to zero, corresponding to a total of 3024 records on the 1st of March.

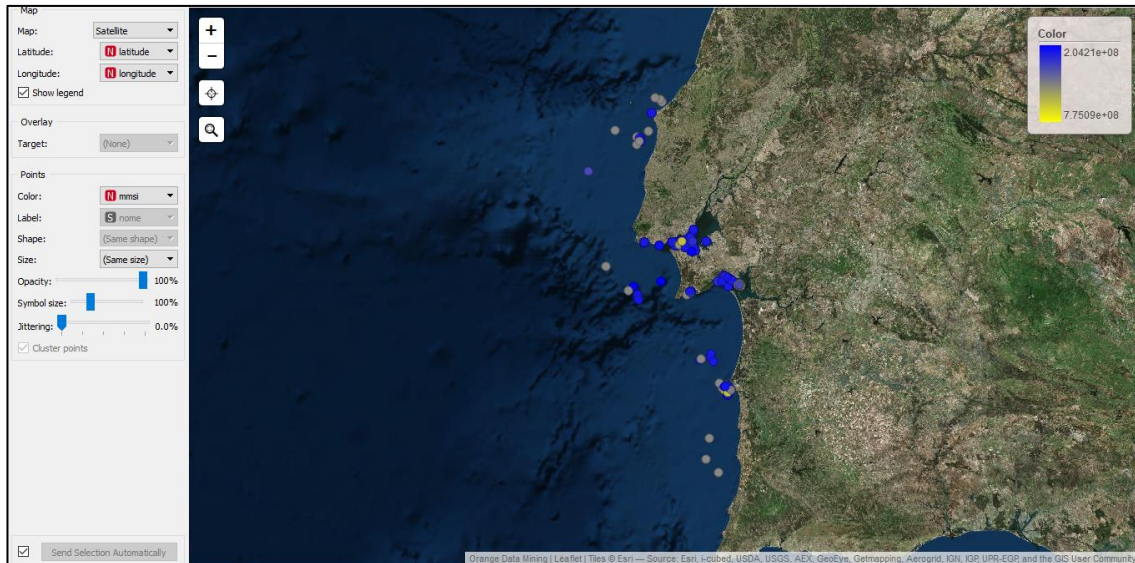


Figure 57 - Records with SOG = 0 on 1st of March, corresponding to a total of 3024 records.

This factor would not be seen as an anomaly if it were not for the distance to the sea bed in those areas, which is far greater than what would be needed for anchoring. There is also the possibility of these being vessels engaged in fishing, which is not unlikely but requires further analysis. Therefore, the next step was to separate records that represented moored vessels from the ones that were at sea. In order to do so, the outlier widget was used, configured as it is represented in Figure 58, after parameters experimentation.

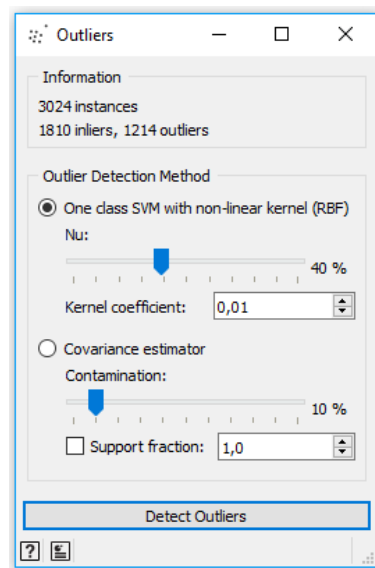


Figure 58 - Outlier widget configuration

With this widget, it was possible to separate the data, resulting in a total of 1810 inliers and 1214 outliers. This is represented in Figure 59, where it is possible to observe the outliers on the left image and the inliers on the right image.

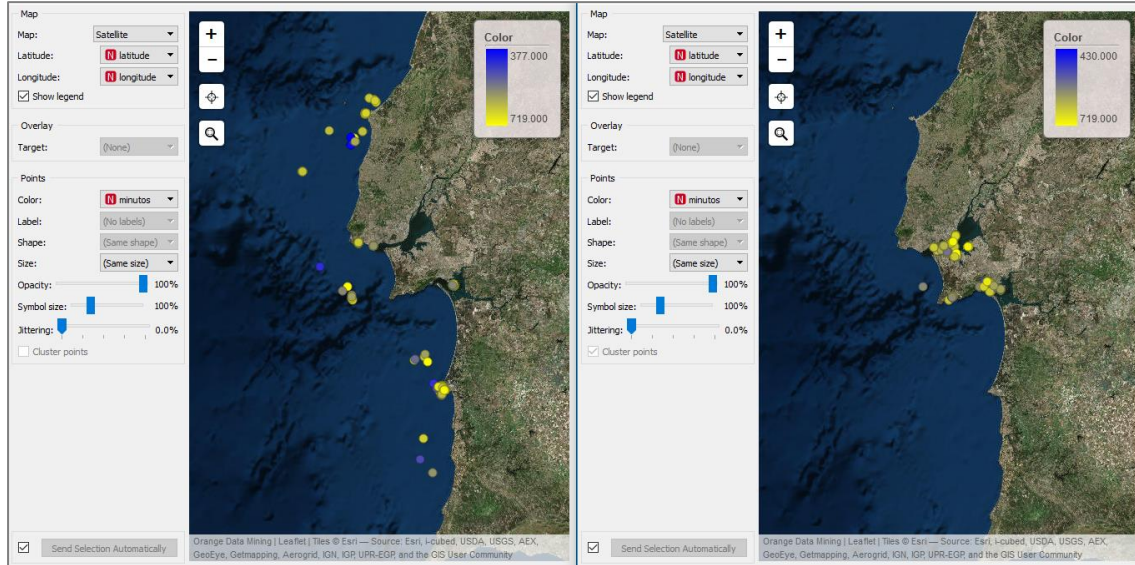


Figure 59 - a) Outlier Records; b) Inlier Records.

From the analysis of Figure 59, it is possible to say that most of the cases where the vessels were at sea were detected in the outliers segment.

Once the records were detected, a second analysis may be done in order to verify if the pattern of movement corresponds to the type of vessel. This analysis is done using the individual segment mentioned above.

In Figure 60, it is represented one of the vessels chosen from the records shown on Figure 59, with a MMSI equal to 255125111. The colours represent different SOGs, being blue SOGs equal to zero and yellow a SOG being equal to nine. Consequently, it is possible to observe one moment where the SOG is equal to zero and another one where the SOG is above 8 knots, showing a pattern that is common with vessels engaged in fishing.

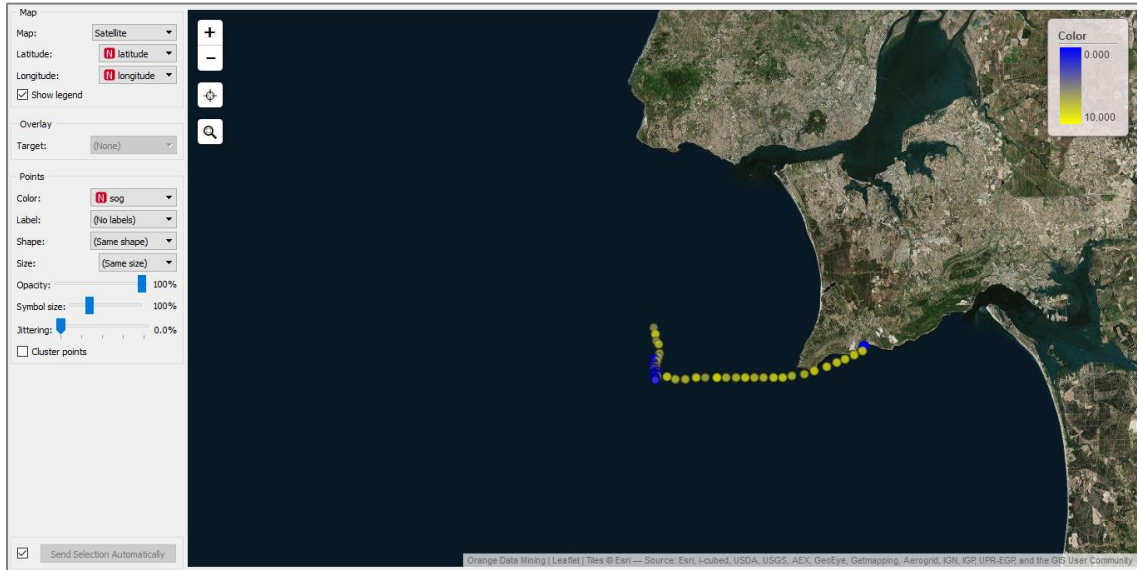


Figure 60 - Vessel with MMSI 255125111 route on 1st of March.

The second branch of the workflow was the SOG outlier detection. In this branch, a simple outlier detection was created, in order to detect records with an anomalous SOG. This branch is represented in Figure 61.

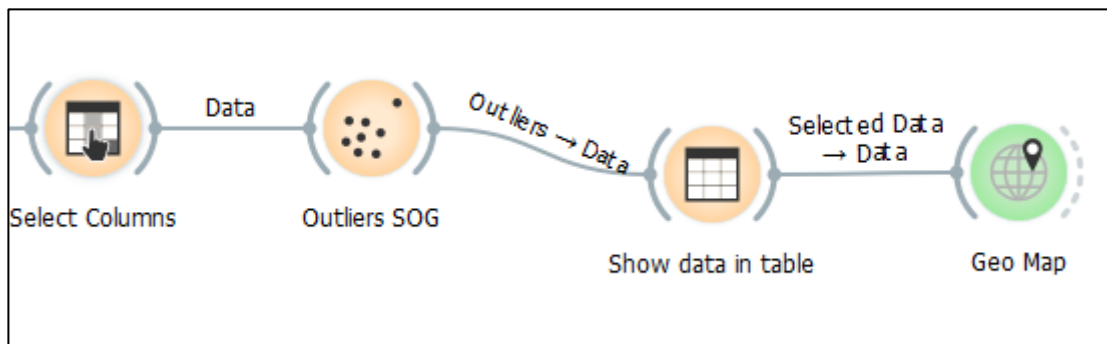


Figure 61 - SOG Outlier detection.

The results are represented in Figure 62. However, due to the long range of SOGs, ranging from 0 to 102 (as it can be observed in the Figure's legend), it is difficult to clarify other records with out of the ordinary SOGs. From these results, it was possible to detect several records with SOGs of 102 knots.

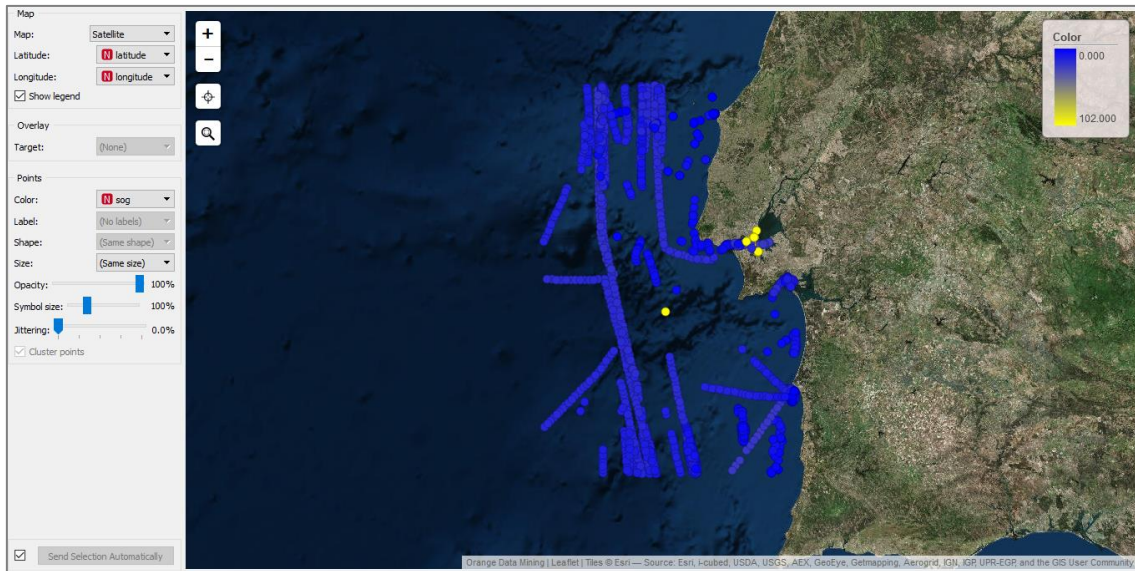


Figure 62 - Results of SOG outlier detection.

Considering that these records are related to vessels, it is highly unlikely that this speed is correct. Upon further research, 102 knots revealed to be the maximum value AIS transponder transmits (Marine Traffic, n.d.). Therefore, it may be concluded that this information results from a system or transmission error.

In order to detect anomalous speeds without system errors, a filter was done, so as to remove speeds over 100 knots. The results are represented in Figure 63.

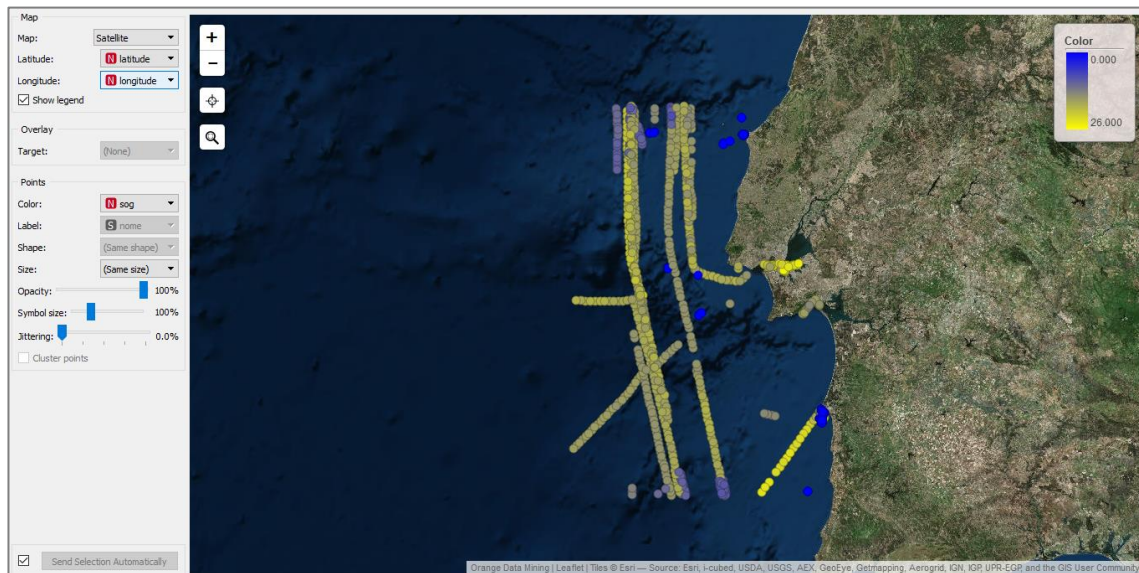


Figure 63 - SOG outlier detection after filtering.

The blue dots represent records with a SOG equal or very close to 0 knots. Upon further analysis, all the records were identified as fishing vessels, as this filter may be indicated to identify this type of ships.

The next branch of this workflow was the Traffic Separation Scheme (TSS) anomaly detection. This branch is represented in Figure 64.

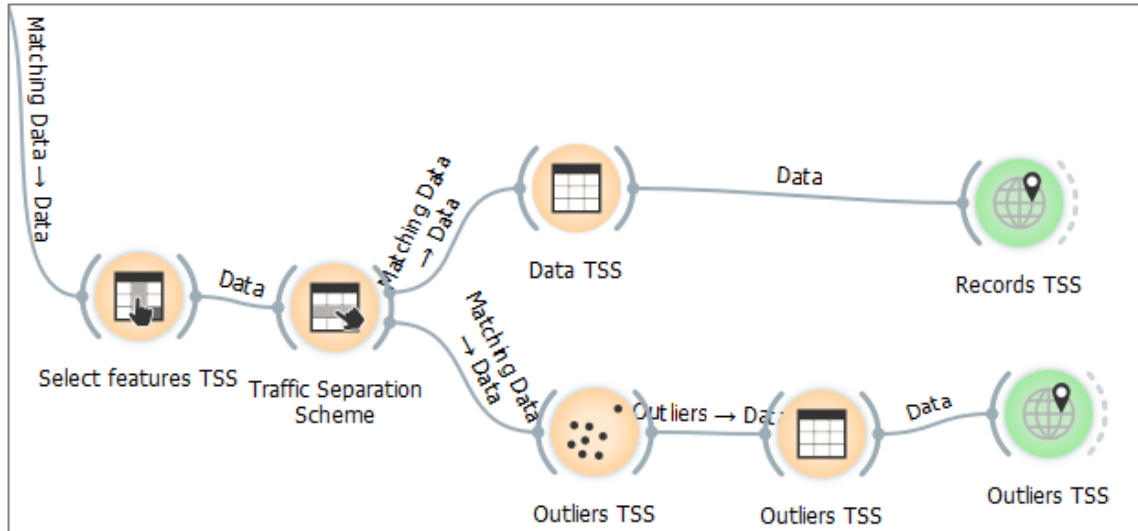


Figure 64 - Traffic separation scheme branch.

For this branch, four features were selected in order to be used in the outliers widget, as it is represented in Figures 65 and 66. The reason these features were selected was to find anomalous records in the TSS, based on an abnormal SOG or COG values in accordance with the vessels' position. A total of 200 outliers were detected in this branch.

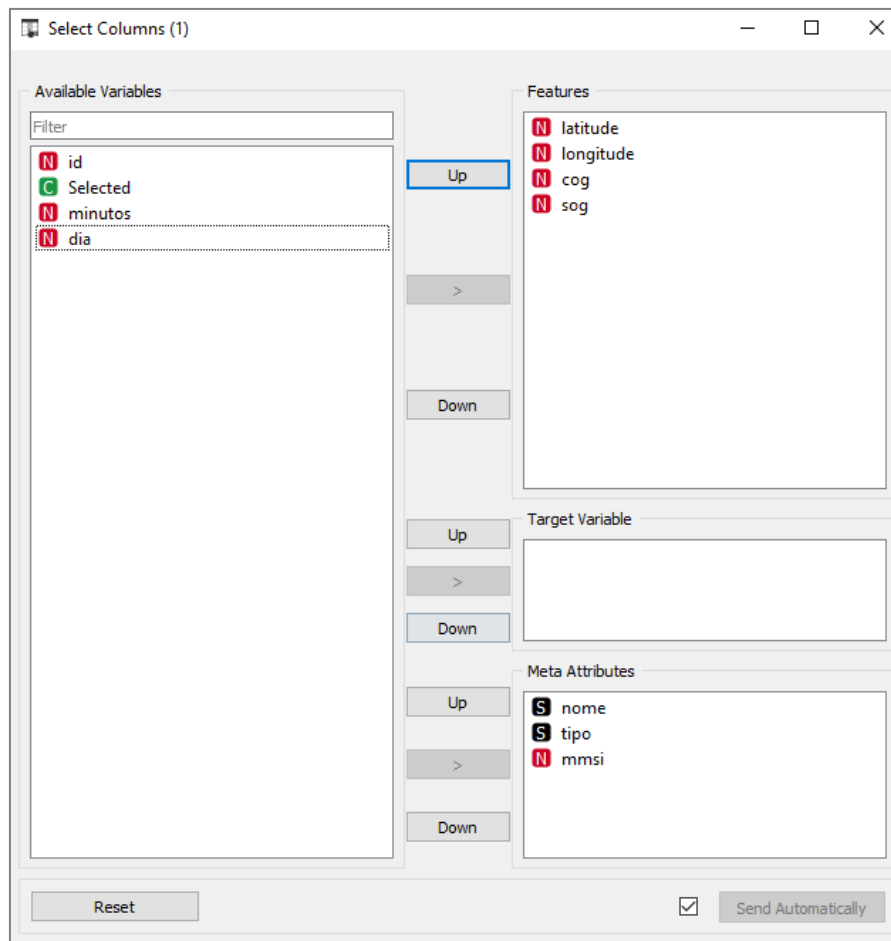


Figure 65 - Features selection for traffic separation scheme.

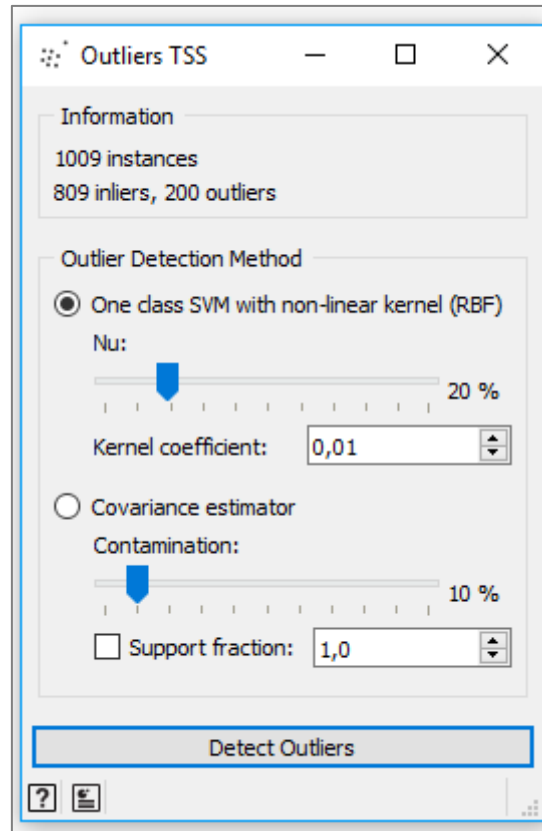


Figure 66 - Outlier widget for traffic separation scheme.

In this branch, only data from the TSS of Cape Roca is selected. This TSS is represented in Figure 67.

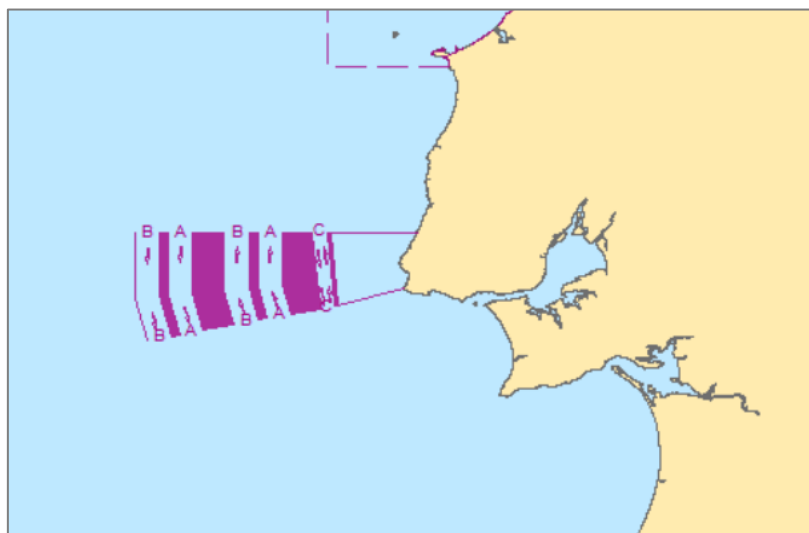


Figure 67 - TSS of Cape Roca.

According to the International Regulations for Preventing Collisions at Sea, there are several rules concerning TSSs (International Rules for Preventing Colisions at Sea,

1972). Therefore, from rule 10 “Traffic separation schemes”, which is presented in Annex B, vessels should “avoid crossing traffic lanes, but if obliged to do so shall cross on a heading as nearly as practicable at right angles to the general direction of traffic flow” (Figure 68).

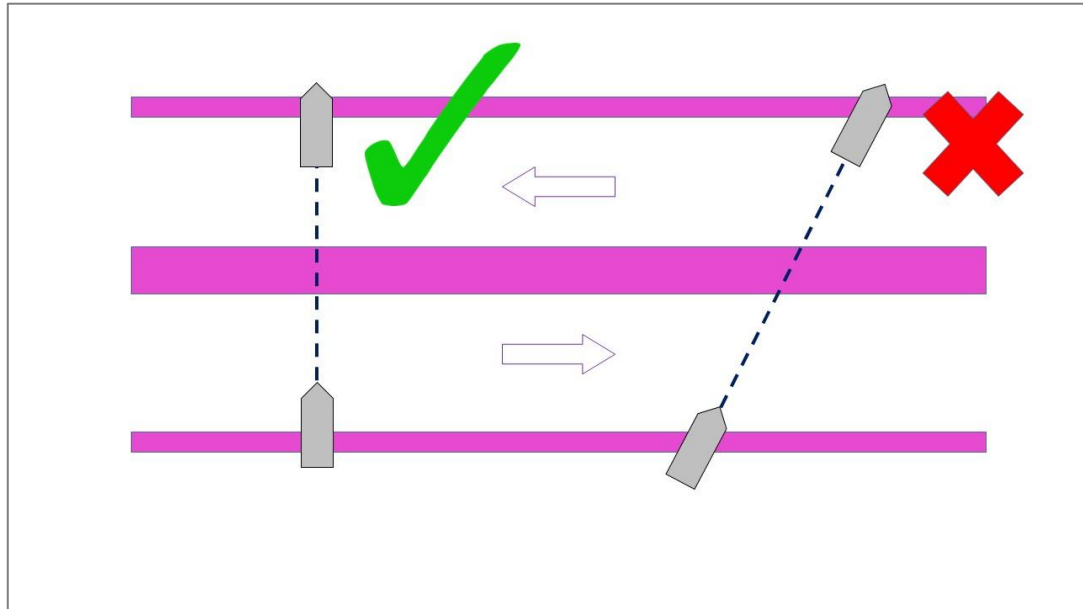


Figure 68 - Crossing a traffic lanes according to International Regulations for Preventing Collisions at Sea.

The results for this branch are presented in Figure 69 a) and b).

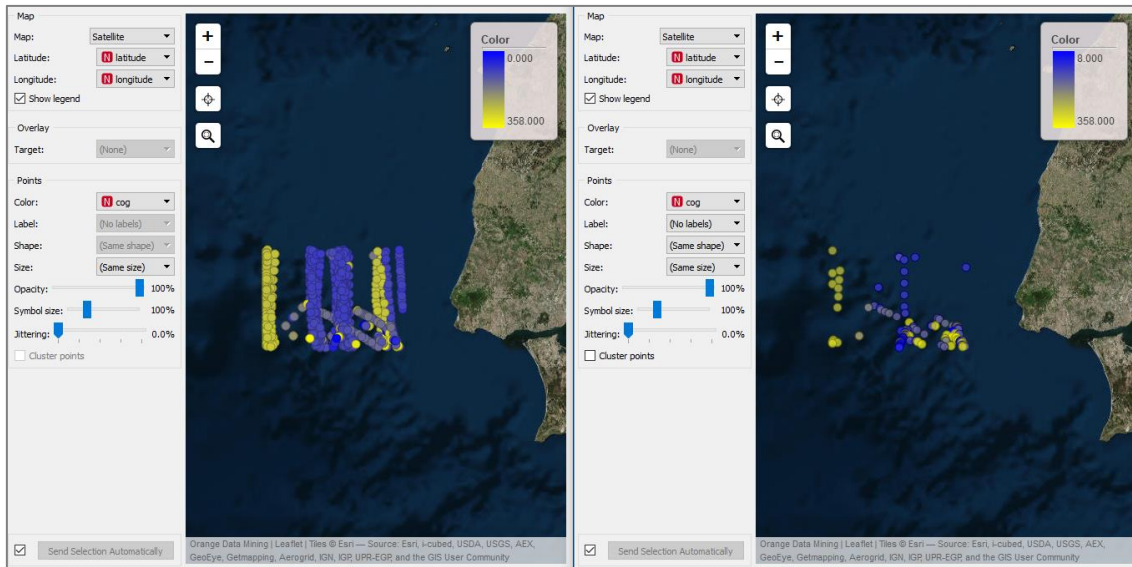


Figure 69 - Traffic separation scheme, a) Every record; b) Outlier records.

The results present two possible anomalous cases, as it is observed in Figure 70.

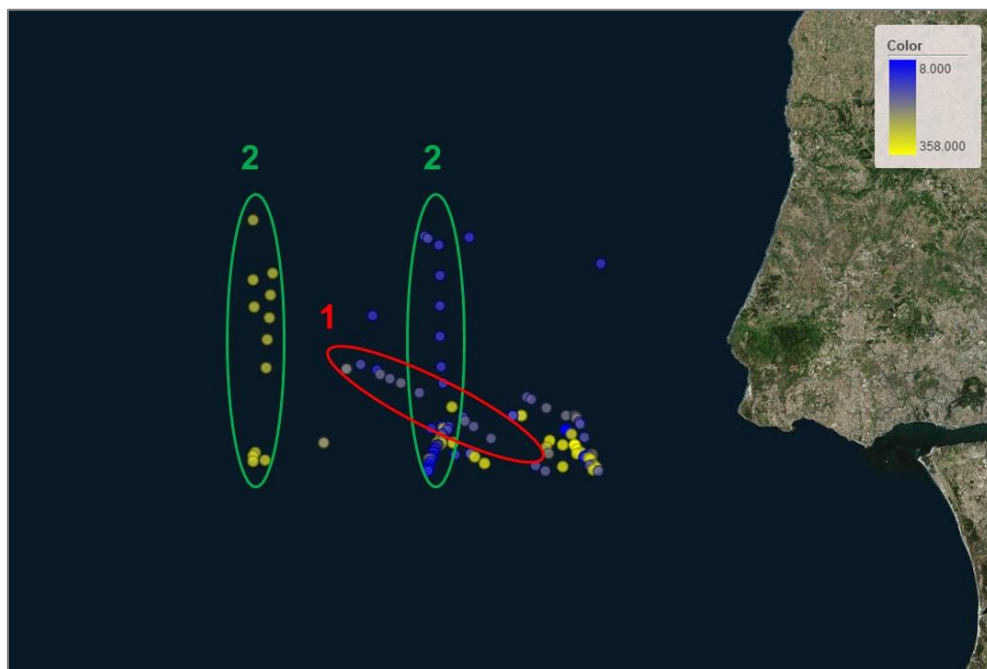


Figure 70 - TSS anomalous records.

In case 1, it is possible to observe a set of records that belong to the same vessel. As it is represented in Figure 70, this vessel is crossing the traffic lanes without doing it at as close as to the normal direction (perpendicular) to the general direction of traffic flow, as it was mentioned above. This practice is not in accordance with the International Regulations for Preventing Collisions at Sea.

On the other hand, in case 2, it is possible to observe two vessels, one in each traffic lane, which according to their positions seem normal. However, their COGs represent abnormal values considering their future positions, as the record from the left corridor has a COG of approximately 265 and the record from the right, a COG of 090. This may mean a system anomaly, as these COG values do not correspond to the ones that are indicated for the vessels to go from one position to the other in that amount of time. The COG values for the left traffic lane should be approximately 180 and for the right traffic lane around 000.

On table 5, it is possible to observe a summary of the Anomaly Detection conducted tests using PostgreSQL database and Orange Data Mining tool.

Table 5 - Summary of the conducted tests.

Tests	Anomaly	Features used	Techniques used	Results
1	Ships with SOGs=0 not moored	Latitude and longitude	One class SVM with non-linear kernel through Outlier widget	The majority of moored vessels were separated. The outliers may be vessels engaged in fishing.
2	Anomalous SOG	Latitude, longitude and SOG	One class SVM with non-linear kernel through Outlier widget	Detected anomalous SOGs of 102 knots, which may indicate a system or transmission anomaly.
3	Anomalous SOGs or COGs in TSS	Latitude, longitude, SOG and COG	One class SVM with non-linear kernel through Outlier widget	Detected anomalous COGs in the TSS and system or transmission anomalies



CHAPTER 6

CONCLUSIONS

- 6.1. Summary
- 6.2. Constraints
- 6.3. Future Work

6. Conclusion

6.1. Summary

This dissertation aims to maximize maritime situational awareness by detecting anomalies in maritime traffic data. The need to have a good maritime situational awareness has been recognized as highly important by every coastal country, more so in a country like Portugal, which has an immense EEZ.

In the first chapter of this dissertation, the topic of this dissertation was introduced. On this chapter it is possible to understand the underlining motivation and relevance of this topic, as well as its main goals.

On Chapter two, a literature review has been done, focusing on maritime situational awareness on the first sub-chapter, moving on to the concept of anomaly and anomaly detection. This chapter ends with a brief review of several data mining tools.

On the third Chapter, the investigation workflow was introduced, as well as an explanation about the origin of the data used in this dissertation and all the steps done in order to include the data in a PostgreSQL database.

On Chapter four, an area of interest was defined and the data mining tool that would be used to analyze the data, as well as the specific functions used in this dissertation.

On the last chapter before conclusions, "Tests and results" the results obtained by using Orange Data Mining Tool are represented. It was on this chapter that it was possible to observe that correlation between data sources can be done through Orange, as well as Outlier Detection in a traffic separation scheme, and other cases with anomalous SOGs and COGs.

A good recognized maritime picture is maintained by using data from different sources. If it is possible to correlate data from different sources, not only is the problem of the overwhelming amount of data somewhat solved, but it allows the detection of different type of anomalies.

Therefore, the goals of this dissertation were for the most part achieved. The results provided by Orange Data Mining turned out to be satisfactory, as it was possible to obtain relevant results, through the detection of anomalies in a simple and easy way, as this tool is very user-friendly. These results proved to be even more satisfactory, as

Orange is in constant development by the community, having a tremendous potential that can be explored even further.

The MARISA project proved to be pertinent, as it contributed for the initial identification of requisites and definition of anomalies, demonstrating to be an influence for this dissertation.

6.2. Constraints

During this dissertation there were some constraints that made it difficult to carry out this research.

The first was the difficulty on obtaining data. The data presented in this dissertation was obtained belatedly, mostly due to bureaucracies, that considering the type of data, and the entities that possess them, are totally understandable. However, taking into consideration the time available for this master thesis, it was a serious limitation. Therefore, a great amount of the available time for the dissertation was used to obtain data and creating the database.

The second one, was the resources available for executing this dissertation. The main factor for this was the lack of a computer that could process the immense amount of data for PostgreSQL database and for Orange analysis. This constraint resulted on the usage of a personal computer, without the ideal processing capabilities that would be desired. As a consequence, the data sets analyzed on Orange had to be, for the most part, reduced.

6.3. Future work

There are several projects that can be done as future work:

- Since the database is already created, it is suggested the addition of meteorological and oceanographic data, and its integration on the analysis;
- The combination of SADAP with Orange is also proposed considering that through SADAP it is possible to evaluate the probability of finding vessels performing infractions. This first assessment would allow to focus Orange's analysis to those specific geographical areas and time periods and judge if the results obtained can identify the reported transgressions. It would also allow a more detailed individual analysis of vessels by showing any existing past infractions;



Data mining for anomaly detection in maritime traffic data

- Orange data mining tool's constant development makes it very important to continue to explore it and its add-ons, such as tools for text analysis from social networks or journals;
- As all the data in the PostgreSQL database is georeferenced, another suggestion would be to explore the variety of spatial operations that PostGIS has to offer.

References

- ARCGIS (n.d.), *Mapping and visualization in ArcGIS for Desktop*, in <http://desktop.arcgis.com/en/arcmap/10.3/main/map/mapping-and-visualization-in-arcgis-for-desktop.htm>, accessed on October 2017.
- (n.d.), *ArcGIS Shapefiles*, in <https://doc.arcgis.com/en/arcgis-online/reference/shapefiles.htm>, accessed on January 2018.
- AHLEMEYER, A., Coleman, S. (2014), *A Practical Guide to Datamining for Business and Industry*, 1st ed., Pondicherry, Wiley.
- ARGUEDAS, V. F., Mazzarella, F., and Vespe, M. (2015), "Spatio-temporal Data Mining for Maritime Situational Awareness", in MTS/IEEE OCEANS 2015 - Genova, Italy, pp.1-8.
- BENNETT, J. (2018), *Orange Data Mining*, in <https://www.predictiveanalyticstoday.com/orange-data-mining/>, accessed on March 2018.
- BHARGAVA, S. C. (2012), *Electrical Measurement Instruments and Measurements*, CRC Press.
- Boost (n.d.), *Normal (Gaussian) Distribution*, in https://www.boost.org/doc/libs/1_38_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/normal_dist.html, accessed on February 2018.
- CAROLAS, Pedro Miguel da Encarnação (2016), *Vigilância e monitorização dos espaços marítimos sob soberania ou jurisdição portuguesa*, Master Thesis in Portuguese Naval Academy, Lisbon.
- CHANDOLA, V., Banerjee, A., Kumar, V. (2009), "Anomaly Detection: A Survey", in ACM Computing Surveys, vol. 41.
- CHEFE DO ESTADO MAIOR DA ARMADA (2016), *Regulamento Interno da Direção de Tecnologias de Informação e Comunicações*, Dispatch nº 50/2016, 10th of May 2016.
- CHEN, C. H. *et al.* (2013), "Knowledge discovery using genetic algorithm for maritime situational awareness", in Elsevier – Expert Systems with Applications.
- DAFTLOGIC (n.d.), *Map area calculator*, in <https://www.daftlogic.com/projects-google-maps-area-calculator-tool.html>, accessed on February 2018.
- DATABASE GUIDE (n.d.), *What is PostgreSQL*, in <https://database.guide/what-is-postgresql/>, accessed on March 2018.
- DIREÇÃO GERAL DE RECURSOS NATURAIS, SEGURANÇA E SERVIÇOS MARÍTIMOS (n.d.), *Zonas Marítimas sob Soberania e ou Jurisdição Portuguesa*, in



<https://www.dgrm.mm.gov.pt/am-ec-zonas-maritimas-sob-jurisdicao-ou-soberania-nacional>, accessed on January 2018.

-
- (n.d.), MONICAP, in <https://www.dgrm.mm.gov.pt/pesca-fisc-MONICAP>, accessed on January 2018.
- DICTIONARY (2018), *Anomaly*, in <http://www.dictionary.com/browse/anomaly>, accessed on February 2018.
- DOREL, P.G. (2013), "European Maritime Safety Agency", *Constanta Maritime University Annals*, vol. 20, Constanta, pp. 271-274.
- ESSENTIAL SQL (n.d.), *What is ACID (atomicity, consistency, isolation, and durability)?*, in <https://www.essentialsql.com/what-is-meant-by-acid/>, accessed on March 2018.
- ESRI (1997), *Esri Shapefile Technical Support*, in <https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>, accessed on February 2018.
- ESTADO MAIOR DA ARMADA (2012), IOA 114 – Conceito de Conhecimento Situacional Marítimo, Marinha Portuguesa.
- EUROPEAN ECONOMIC COMMUNITY (2013), Regulation EEC 2847/93, European Communities Official Paper, Series L-261, October 12th 1993.
- _____ (2018), *Ata da Sessão Ordinária do Grupo de Trabalho para o Conhecimento Situacional Marítimo (GT-CSM)*, December 7th 2017, pp. 1-14.
- EUROPEAN MARITIME SAFETY AGENCY (n.d.), *LRIT Cooperative Data Centre*, in <http://www.emsa.europa.eu/lrit-main/lrit-home.html>, accessed on February 2018.
- FAYYAD, U., Piatetsky-Shapiro, G., Smyth, P. (1996), "From Data Mining to Knowledge Discovery in Databases", in *AI Magazine*, vol. 17, pp. 1-18.
- GERMAIN, E. (1997), "The Coming Revolution in NATO Maritime Command and Control", in The MITRE Corporation.
- GISGEOGRAPHY (2018), *World Geodetic System (WGS84)*, in <https://gisgeography.com/wgs84-world-geodetic-system/>, accessed on February 2018.
- GONZALEZ, F., Dasgupta, D., Kozma, R. (2002), "Combining negative selection and classification techniques for anomaly detection", in *IEEE Congress on Evolutionary Computation*, Honolulu, USA.
- HAN, J., Kamber, M., Pei, J. (2012), *Data Mining Concepts and Techniques*, 3rd ed., Waltham, Elsevier.
- HERNANDEZ-CASTRO, J., Roberts, D. L. (2015), "Automatic detection of potentially illegal online sales of elephant ivory via data mining", in *PeerJ Computer Science*.



- INOV (n.d.), *MONICAP*, in <http://www.inov.pt/index/casos-de-sucesso/165-MONICAP.html>, accessed on February 2018.
- INSTITUTO HIDROGRÁFICO (n.d.), *Bluemassmed*, in <http://www.hidrografico.pt/bluemassmed-ih.php>, accessed on January 2018.
- INTERNATIONAL MARITIME ORGANIZATION (n.d.), *AIS transponders*, in <http://www.imo.org/en/OurWork/Safety/Navigation/Pages/AIS.aspx>, accessed on February 2018.
- JORNAL DA ECONOMIA DO MAR (n.d.), *O quebra-cabeças do shipping*, <http://www.jornaldaeconomiadomar.com/o-quebra-cabecas-do-shipping/>, accessed on February 2018.
- KANTARDZIC, M. (2011), *Data Mining Concepts, Models, Methods and Algorithms*, 2nd ed., New Jersey, John Wiley & Sons.
- KOWALCZYK, R. (2009), *Service-Oriented Computing: Agents, Semantics, and Engineering*, 1st ed., Budapest, Springer.
- MAO, S. et al. (2016), *An Automatic Identification System (AIS) Database for Maritime Trajectory Prediction and Data Mining*, Proceedings in Adaptation, Springer.
- Marine Traffic (n.d.), *Automatic Identification System*, in <https://www.marinetraffic.com/en/ais/home/centerx:-47.0/centery:30.8/zoom:2>, accessed on March 2018.
- (n.d.), *What is the typical range of AIS?*, in <https://help.marinetraffic.com/hc/en-us/articles/203990918--What-is-the-typical-range-of-the-AIS->, accessed on February 2018.
- (n.d.), *Historical AIS data*, in <https://www.marinetraffic.com/en/p/ais-historical-data>, accessed on February 2018.
- MARINHA (2018), *Operações Marítimas*, in <http://www.marinha.pt/pt-pt/meios-operacoes/comando-apoio/centros/Paginas/Operacoes-Maritimas.aspx>, accessed on February 2018.
- MARITIME SAFETY AND SECURITY INFORMATION SYSTEM (2008), *MSSIS*, in <https://mssis.volpe.dot.gov/Main/>, accessed on January 2018.
- MARTINEAU, E., Roy, J. (2011), *Maritime Anomaly Detection: Domain Introduction and Review of Selected Literature*, Defence Research and Development Canada, Valcartier.
- MCANET (n.d.), *Automatic Identification Systems*, in https://mcanet.mcga.gov.uk/public/c4/solas/solas_v/Annexes/Annex17.htm#ais, accessed on January 2018.
- MELO, Hugo Daniel Almeida de (2011), *Módulo de análise do quadro situacional marítimo para apoio a missões de vigilância e fiscalização marítima*, Master Thesis in Portuguese Naval Academy, Lisbon.



- MICROSOFT (n.d.), *Excel Specifications and Limits*, in <https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3>, accessed on October 2017.
- MITCHELL, P. (2013), *Network Centric Warfare and Coalition Operations*, 1st ed., London, Routledge.
- MOREIRA, A. (2013), "Synthetic Aperture Radar (SAR): Principles and Applications", in Advanced Training Course in Land Remote Sensing, Athens.
- NATIONAL RESEARCH COUNCIL (2008), *Maritime Security Partnerships*, Washington, The National Academy Press.
- NAVIGATION CENTER (n.d.), *How AIS Works*, in <https://www.navcen.uscg.gov/?pageName=AISworks>, accessed on January 2018.
- _____ (n.d.), How does AIS compare and contrast with VMS?, in https://www.navcen.uscg.gov/pdf/AIS/Q_AIS_vs_VMS_Comparison.pdf, accessed on February 2018.
- NORTH AMERICAN JOB BANK INTERNATIONAL NETWORKING (n.d.), *Technical Project Mgr Baseline for Rapid Iterative Transformational Experimentation*, in <http://www.najobbank.com/Technical-Project-Mgr-Baseline-for-Rapid-Iterative-Transformational-Experimentation-1044-1-123789.html>, accessed on January 2018.
- ORACLE (n.d.), *Data mining concepts*, in https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHD FGCIJ, accessed on February 2018.
- _____ (n.d.), *Orange widget catalog*, in <https://orange.biolab.si/toolbox/>, accessed on March 2018.
- PEREIRA, Luís Carlos de Sousa (2010), *Maritime Situational Awareness. Portugal e o efectivo controlo do espaço estratégico de interesse nacional permanente em ambiente marítimo tendo em conta as novas soluções preconizadas no conceito MSA*, Instituto Superior de Estudos Militares, Lisboa.
- PORDATA (2016), *Portuguese surface*, in <https://www.pordata.pt/Europa/Superfície-2489>, accessed on February 2018.
- POSTGIS (n.d.), *PostGIS*, in <https://postgis.net/>, accessed on November 2017.
- POSTGRESQL (n.d.), *About PostgreSQL*, in <https://www.PostgreSQL.org/about/>, accessed on November 2017.
- PREDICTIVE ANALYTICS TODAY (n.d.), *Comparison Dashboard: Orange Data mining vs WEKA Data mining vs RapidMiner*, in <https://www.predictiveanalyticstoday.com/compare/orange-data-mining-vs-weka-data-mining-vs-rapidminer-starter-edition/>, accessed on March 2018.

- REHMAN, A., Mahmood, S. (2009), "WEKA & KNIME Open Source Machine Learning", in Third International Conference on Open-Source Systems and Technologies, Pakistan, www.uet.edu.pk/Conferences/icosst2009/presentations_2009/OSSW_Presentations/4_WEKA_x_KNIME_Open_Source_Machine_Learning_Tools_OSSW_ICO_SST_2009.pdf, accessed on February 2018.
- ROSENBERG, M. (2018), *What is the Distance Between a Degree of Latitude and Longitude*, in <https://www.thoughtco.com/degree-of-latitude-and-longitude-distance-4070616>, accessed on March 2018.
- SAILWX (n.d.), *Sailwx database*, in <https://www.sailwx.info/shiptrack/search.phtml>, accessed on February 2018.
- SANDIA NATIONAL LABORATORIES (n.d.), *What is Synthetic Aperture Radar (SAR)?*, in http://www.sandia.gov/radar/what_is_sar/index.html, accessed on February 2018.
- SCHILLING, D. R. (2013), *Knowledge Doubling Every 12 Months, Soon to be Every 12 Hours*, in <http://www.industrytap.com/knowledge-doubling-every-12-months-soon-to-be-every-12-hours/3950>, accessed on February 2018.
- SEIBERT, M. (2009), "Maritime Anomaly Detection", Workshop on Detection of Anomalous Behaviours in Maritime Environments, Carnegie Mellon University.
- SHAH, D. (2017), *Data Mining Tools*, in <https://towardsdatascience.com/data-mining-tools-f701645e0f4c>, accessed on February 2018.
- SHARDA, R., Delen, D., Turban, E. (2014), *Business Intelligence and Analytics: Systems for Decision Support*, 3rd ed., Pearson, Edinburgh.
- SOARES, C. G. et al. (2012), *Maritime Engineering and Technology*, CRC Press.
- SOUSA, Laura Sofia Neves de (2013), *Indicadores de risco de incidentes marítimos com base em dados do sistema de monitorização contínua das atividades de pesca*, Master Thesis in Portuguese Naval Academy, Lisbon.
- SULEMANE, Yazide Abdul Carimo (2015), *Ferramentas para a Análise dos Padrões de Tráfego da Barra Sul do Porto de Lisboa*, Master Thesis in Portuguese Naval Academy, Lisbon.
- TETREAULT, B. J. (2005), "Use of the Automatic Identification System (AIS) for maritime domain awareness (MDA)", in MTS/IEEE OCEANS 2005 - Washington, pp.1-5.
- TRANSPORTES E NEGÓCIOS (2017), *Lisboa com o melhor resultado em nove anos*, in <https://www.transportesenegocios.pt/lisboa-com-o-melhor-resultado-em-nove-anos/>, accessed on December 2017.
- UNITED NATIONS (2017), *Review of Maritime Transport 2017*, in United Nations Conference on Trade and Development, New York and Geneva, pp.1-21.
- URBANO, F., Cagnacci, F. (2014), *Spatial Database for GPS Wildlife Tracking Data*, 1st ed., London, Springer.



VELOSO, Ricardo José Santos (2015), “A Partilha de Dados no Mar”, Anais do Clube Militar Naval, year CXLV, Lisbon, pp.737-758.

XSEALENCE (n.d.), *Sistemas Xsealence*, in <http://www.xsealence.pt/sistemas-xsealence/>, accessed on February 2018.

————— (n.d.), *Equipamentos de Monitorização de Embarcações de Pesca*, in <http://www.xsealence.pt/portfolio/MONICAP/>, accessed on February 2018.

ZUPAN, B. et al. (2001), “Orange and Decisions-at-Hand: Predictive Data Mining and Decision Support”, pp.1-12.

Appendix A

Matlab code to convert data to .xlsx format:

```
%PASSAR DADOS AIS DE *MAT PARA EXCEL

% Dia 1
load('E:\DADOS AIS\2017\10\F\F01Oct2017.mat');
c=length(footprint_aom);
d=cell(c,11); % cria um conjunto de células (tem que ser células e não
matrizes, por serem dados do tipo string)
for ir=1:c
    a=footprint_aom(ir,1);
    e=footprint_aom(ir,2);
    b=sprintf('%0f',a); % passar MMSI p string / '%0f' significa que
não vai ter nenhuma casa decimal
    f=datestr(e,'dd/mm/yyyy HH:MM'); % converter o datenum no formato
indicado
    d{ir,1}=b;
    d{ir,2}=f(1:2);%dia
    d{ir,3}=f(4:5);%mes
    d{ir,4}=f(7:10);%ano
    h=str2num(f(12:13));%horas
    m=str2num(f(15:16));%minutos
    min = (h*60) + m ;
    d{ir,5}=min;
    d{ir,6}=footprint_aom(ir,3);
    d{ir,7}=footprint_aom(ir,4);
    d{ir,8}=footprint_aom(ir,5);
    d{ir,9}=footprint_aom(ir,6);
    d{ir,10}=footprint_aom(ir,7);
    d{ir,11}=footprint_aom(ir,8);
end

% passar dados para formato xlsx (excel)
xlswrite('AIS_01Oct.xlsx',d);
```



Data mining for anomaly detection in maritime traffic data

Annex A – International Requirements for AIS Carriage

According to Chapter V, Regulation 19.2.4 of SOLAS Convention:

“2.4. All ships of 300 gross tonnage and upwards engaged on international voyages and cargo ships of 500 gross tonnage and upwards not engaged on international voyages and passenger ships irrespective of size shall be fitted with an automatic identification system (AIS), as follows:

2.4.1. Ships constructed on or after 1 July 2002;

2.4.2. Ships engaged on international voyages constructed before 1 July 2002:

2.4.2.1. in the case of passenger ships, not later than 1 July 2003;

2.4.2.2. in the case of tankers, not later than the first survey for safety equipment* on or after 1 July 2003;

* Refer to regulation I/8

2.4.2.3. in the case of ships, other than passenger ships and tankers, of 50,000 gross tonnage and upwards, not later than 1 July 2004;

2.4.2.4. in the case of ships, other than passenger ships and tankers, of 300 gross tonnage and upwards but less than 50,000 gross tonnage, not later than the first safety survey after 1 July 2004 or by 31 December 2004, whichever occurs earlier; and

2.4.3. ships not engaged on international voyages constructed before 1 July 2002, not later than 1 July 2008;

2.4.4. the Administration may exempt ships from the application of the requirements of this paragraph when such ships will be taken permanently out of service within two years after the implementation date specified in subparagraphs .2 and .3”.



Data mining for anomaly detection in maritime traffic data

Annex B - Rule 10: Traffic Separation Schemes

According to the International Regulations for Preventing Collisions at Sea – Rule 10:

(a) This Rule applies to traffic separation schemes adopted by the Organization and does not relieve any vessel of her obligation under any other rule.

(b) A vessel using a traffic separation scheme shall:

(i) proceed in the appropriate traffic lane in the general direction of traffic flow for that lane;

(ii) so far as practicable keep clear of a traffic separation line or separation zone;

(iii) normally join or leave a traffic lane at the termination of the lane, but when joining or leaving from either side shall do so at as small an angle to the general direction of traffic flow as practicable.

(c) A vessel shall, so far as practicable, avoid crossing traffic lanes but if obliged to do so shall cross on a heading as nearly as practicable at right angles to the general direction of traffic flow.

(d)

(i) A vessel shall not use an inshore traffic zone when she can safely use the appropriate traffic lane within the adjacent traffic separation scheme. However, vessels of less than 20 metres in length, sailing vessels and vessels engaged in fishing may use the inshore traffic zone.

(ii) Notwithstanding subparagraph (d)(i), a vessel may use an inshore traffic zone when en route to or from a port, offshore installation or structure, pilot station or any other place situated within the inshore traffic zone, or to avoid immediate danger.

(e) A vessel other than a crossing vessel or a vessel joining or leaving a lane shall not normally enter a separation zone or cross a separation line except:

(i) in cases of emergency to avoid immediate danger;

(ii) to engage in fishing within a separation zone.

(f) A vessel navigating in areas near the terminations of traffic separation schemes shall do so with particular caution.



- (g) A vessel shall so far as practicable avoid anchoring in a traffic separation scheme or in areas near its terminations.
- (h) A vessel not using a traffic separation scheme shall avoid it by as wide a margin as is practicable.
- (i) A vessel engaged in fishing shall not impede the passage of any vessel following a traffic lane.
- (j) A vessel of less than 20 metres in length or a sailing vessel shall not impede the safe passage of a power-driven vessel following a traffic lane.
- (k) A vessel restricted in her ability to manoeuvre when engaged in an operation for the maintenance of safety of navigation in a traffic separation scheme is exempted from complying with this Rule to the extent necessary to carry out the operation.
- (l) A vessel restricted in her ability to manoeuvre when engaged in an operation for the laying, servicing or picking up of a submarine cable, within a traffic separation scheme, is exempted from complying with this Rule to the extent.



Data mining for anomaly detection in maritime traffic data
